

Высшее профессиональное образование

Ю. Н. Марчук

МОДЕЛИ ПЕРЕВОДА

Учебное пособие



Языкознание

ВЫСШЕЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАНИЕ

Ю. Н. МАРЧУК

МОДЕЛИ ПЕРЕВОДА

*Учебное пособие
для студентов учреждений
высшего профессионального образования*



Москва
Издательский центр «Академия»
2010

УДК 655.525.3/4(075.8)
ББК 81.2-7я73
М30

Рецензенты:
доктор филологических наук,
профессор кафедры английской филологии
Российского государственного социального университета *К. Я. Авербух*;
доктор филологических наук, профессор кафедры иностранных языков № 3
Российского университета дружбы народов *А. Л. Семенов*

Марчук Ю. Н.

М30 Модели перевода : учеб. пособие для студ. учреждений высш. проф. образования / Ю. Н. Марчук. — М.: Издательский центр «Академия», 2010. — 176 с.

ISBN 978-5-7695-6991-3

В учебном пособии излагаются вопросы моделирования перевода с одного естественного языка на другой. Рассматриваются проблемы построения моделей как «человеческого», так и машинного перевода. Основное внимание уделяется формальному моделированию перевода и новым информационным технологиям в их приложении к лингвистическим задачам. Анализируются современные системы машинного перевода, их структура и эффективность.

Для студентов учреждений высшего профессионального образования. Может быть полезно всем, интересующимся проблемами автоматического анализа и синтеза естественно-языковых текстов в новых условиях языковой коммуникации.

УДК 655.525.3/4(075.8)
ББК 81.2-7я73

*Оригинал-макет данного издания является собственностью
Издательского центра «Академия», и его воспроизведение любым способом
без согласия правообладателя запрещается*

© Марчук Ю.Н., 2010
© Образовательно-издательский центр «Академия», 2010
ISBN 978-5-7695-6991-3 © Оформление. Издательский центр «Академия», 2010

ПРЕДИСЛОВИЕ

Учебное пособие составлено на основе лекций, которые в течение многих лет его автор читал в различных вузах России и за рубежом в рамках курсов теории перевода и переводческой практики, машинного перевода, прикладной лингвистики и других лингвистических дисциплин, центром которых был перевод как один из важных и чрезвычайно сложных видов языковой деятельности человека. Нельзя сказать, что теории перевода и переводческой практике сейчас не уделяется внимания. Скорее, наоборот: современная языковая ситуация и мода на изучение иностранного языка (в основном английского), вызванная к жизни реформами нынешнего русского общества, породили множество практических руководств по изучению иностранного языка, теоретических и прикладных работ по теории и практике перевода, словарей и пособий, разного рода методов и средств ускоренного обучения языкам и т.д. Молодежь достаточно быстро овладевает основами переводческого дела. Однако перевод является настолько сложным и своеобразным видом умственной и языковой деятельности человека, настолько разносторонним явлением, что подходов к его изучению и соответствующих исследований и руководств никогда не будет слишком много. Тем более что чаще всего эти исследования затрагивают отдельные вопросы переводческой деятельности.

Перевод в настоящее время является чрезвычайно актуальным видом языковой деятельности в связи с тем, что постоянно расширяется объем массовой коммуникации, в общение вступают все новые языки, и единственным средством эффективного преодоления языковых барьеров является перевод. Языковые барьеры являются главным препятствием для распространения информации, основной объем которой относится к научно-технической, определяющей технический прогресс и развитие цивилизации. Существуют и другие средства преодоления языковых барьеров, а именно единый язык и обучение иностранным языкам на разных уровнях образования (далее мы рассмотрим конкретно вопрос об искусственных универсальных языках). Изучение же иностранных языков в средней школе и вузах, как показали исследования, проведенные информационными организациями по проверке эффективности изучения иностранных языков, за редкими исключениями, не является достаточным для эффективного преодоления языковых барьеров.

В современную эпоху глобализации естественным является вопрос о том, не будет ли все мировое сообщество, особенно его наиболее образованный слой, общаться на английском языке (в его американском варианте), поскольку США выступают в роли ведущей мировой державы, и к этому есть определенные технические и коммуникационные основания, например Интернет, не говоря уже о социальных и политических тенденциях. Предполагается, однако, что на этот вопрос следует ответить отрицательно. Каждая страна (нация, народность, племя) желает говорить на том языке, который является для нее родным. Количество языков в мире составляет несколько тысяч — подсчет зависит от того, как велика языковая общность, объединенная языком, насколько язык отличается от диалекта и т. п., в любом случае несколько тысяч языков реально существуют. Одни из них отмирают в связи с уходом из жизни говорящих на них людей, зато есть факты возрождения некоторых из них. Так, в Великобритании восстановлен язык одного из кельтских племен, в свое время весьма многочисленных и сопротивлявшихся нашествию норманнов. Последний носитель этого языка ушел из жизни еще в XVII в. На острове Тенериф (Канарские острова) живут два человека, знающие язык свиста, возникший в пору римского владычества в Африке. В России сделано немало для поддержки и сохранения языков народов Крайнего Севера и Дальнего Востока, а также других регионов страны, где проживали небольшие группы людей, говоривших на древних языках и диалектах. Так что вряд ли приходится говорить о том, что процесс отмирания языков и/или глобализация приведут к одному естественному языку в мире как универсальному средству общения.

В связи с этим перевод является и остается единственным эффективным средством преодоления языковых барьеров. Специальные исследования, посвященные вопросу «что переводится», показывают, что до 80—90 % переводимых текстов — научные или технические. Далее следуют юридические, политические, учебные и прочие. Художественный перевод, по подсчетам Европейской комиссии, занимает в общем объеме переводов 0,3 %. Научно-технический перевод вносит значительный вклад в научно-технический и социальный прогресс человечества. При этом, однако, «человеческий» перевод (human translation) занимает много времени. Так, например, по данным Всесоюзного центра переводов научно-технической литературы и документации (ВЦП), крупнейшей переводческой организации СССР (теперь — Всероссийский центр переводов), нормой научно-технического перевода считалось время 10 дней на авторский лист (24 страницы машинописного текста). Вряд ли эта цифра уменьшилась сегодня, поскольку ресурсы умственной деятельности человека ограничены, а перевод является высокоинтеллектуальной разновидностью такой деятельности.

Кроме того, «человеческий» перевод достаточно дорог с экономической точки зрения, поскольку требует оплаты труда переводчиков и других услуг, связанных с осуществлением и представлением переводов, например получение уже выполненных работ из некоего специального фонда в режиме координации переводов.

Кстати сказать, в русском языке нет хорошего эквивалента английскому *human translation*, поскольку слово «человеческий» имеет много других коннотаций, а термин «ручной» не очень удачен, поскольку перевод выполняется не руками, а головой. Можно заметить, что и в английском есть такие случаи: так, для слова «русский» и «россиянин» существует один эквивалент *Russian*, который не передает различия в национальности — «россиянином» может быть и татарин или чукча.

Идея применить компьютер к переводу не нова и к настоящему времени насчитывает более пятидесяти лет. Она зародилась практически с первым появлением электронных вычислительных машин — ЭВМ, как раньше называли компьютеры. Одним из первых применений ЭВМ было кодирование и декодирование сообщений. Разработчики первых ЭВМ высказали идею о применении их к переводу по следующему рабочему соображению: представим себе, что один естественный язык является кодом другого. Если мы расшифруем этот код, то вот вам и перевод с одного языка на другой. Таким образом, язык приравнялся к некоторому коду. Можно в связи с этим говорить о том, что первой моделью языка была модель некоторого кода (об этом будет сказано позже). Такая модель, однако, не смогла решить проблему перевода с помощью ЭВМ, главным образом, потому, что единицы языкового кода, если можно так называть лингвистические сущности, многозначны в отличие от кодовых единиц. Тем не менее такой взгляд на язык сыграл большую роль не только в теории перевода, но и в прикладном и теоретическом языкознании в целом.

Мы живем в эпоху новых информационных технологий, которые в первую очередь связаны с возросшими возможностями компьютера, с проникновением информатики в нашу повседневную жизнь. Машина еще не умеет мыслить, однако ее возможности увеличиваются, и рано или поздно некоторые функции человеческого интеллекта могут быть переданы компьютеру. Академик Ю. В. Рождественский совершенно справедливо высказал мысль о том, что машинный перевод является центральной проблемой искусственного интеллекта. Действительно, если мы представим себе, что перевод с одного языка на другой не есть простая замена слов одного языка на слова другого, а требует проникновения в смысл, содержание, учета многих интеллектуальных факторов, таких, как языковые картины мира отправителя и получателя перевода, сопоставления семантики языков, учета внешних факторов общения и прочих обстоятельств перевода, сложного взаимодействия разных языковых уровней, то связь

перевода с мыслительной деятельностью, с интеллектом никак не может казаться парадоксальной. Связь языка с естественным интеллектом специально изучается, поскольку «язык и интеллект человека неразрывно связаны и потому представляют собой важнейший источник сведений друг о друге» [Рябцева 2005*].

В свете всего этого модели перевода безусловно имеют большое значение как для прикладной лингвистики, прежде всего совершенствования лингвистических технологий, так и для теоретической, поскольку перевод есть одна из высших форм интеллектуальной деятельности человека.

Теория перевода возникла сравнительно недавно, в 30-х гг. XX в., хотя вопросами и практикой перевода лингвисты интересовались давно: переводы религиозных текстов, например Библии, известны с древних времен. Как составляющая прикладного языкознания, теория перевода возникла в СССР в результате осмысления переводческой деятельности как самостоятельного занятия в рамках общелингвистических теорий. Модели перевода, результат анализа переводческой деятельности, достаточно разнообразны, поскольку создавались с разных теоретических и практических позиций и с разными целями — обучения переводу, связи его с различными уровнями языка, отношениями с экстралингвистическими факторами, отношением перевода к передаче смысла и содержания, автоматизации перевода и т. п. Автор настоящего пособия пытается дать обзор и анализ основных направлений и результатов моделирования перевода, обращая главное внимание не столько на философские и абстрактные аспекты, сколько на возможности четкого, достаточно формального, допускающего проверку описания основных черт моделирования перевода.

Суммируя все вышесказанное, можно задать вопрос: чем отличается данная книга от многочисленных работ по теории и практике перевода, появившихся в настоящее время? Тем, что в ней главное внимание уделено формализации процесса перевода, чтобы как можно большую часть работы передать новым информационным технологиям. Формализация в данном случае понимается не как представление переводческих проблем в виде математических формул, а представление (или поиски представления) аспектов лингвистической стороны в переводе, в том числе и уровня семантики, в виде некоторых непротиворечивых языковых формализмов. Лингвистическая сторона новых информационных технологий представляется автору гораздо более важной, чем сторона формально-математическая, поэтому ей уделяется особое внимание. Однако и последняя также будет рассмотрена в настоящем пособии. Такой подход нечасто применяется в современных теоретических и практических разработках. Надеюсь, что читатель найдет для себя что-то интересное и полезное в настоящем изложении.

* Здесь и далее ссылки в скобках см. Список литературы, с. 168.

РАЗДЕЛ I

СОВРЕМЕННЫЕ ПРОБЛЕМЫ ЯЗЫКОВОГО ПЕРЕВОДА

Глава 1

МЕСТО ЯЗЫКОВОГО ПЕРЕВОДА В СОВРЕМЕННОЙ КОММУНИКАЦИИ

Преодоление языковых барьеров

Век информатизации и компьютеризации значительно расширил возможности общения в человеческом обществе. Интернет и другие коммуникационные линии сделали возможным установление связей в реальном времени практически каждого грамотного жителя Земли с любым другим грамотным человеком. Распространение информации, в первую очередь научно-технической, но также и всякой другой — политической, юридической, текущей, социальной и пр. — решающим образом определяет прогресс человечества и каждой нации в цивилизованном обществе. На пути информации есть и будут барьеры — политические, географические, экономические (стоимость коммуникации), таможенные, коммерческие (сохранение коммерческой тайны), военные (сохранение государственных секретов) и другие. Однако самый главный из них — языковой. Народы мира говорят на разных языках. Чем более продвинута данная нация на пути цивилизации и экономического развития, тем большую цену она готова платить за право общаться на своем языке. Такую формулу предлагают современные руководители информационных служб, обзревая состояние рынка информационных услуг и сферу коммуникации [Better Translation for Better Communication 1983].

Есть **три способа преодоления языковых барьеров**. Первый — использование единого, одного языка для всех жителей Земли — потенциальных коммуникантов. В этом способе, в свою очередь, есть две разновидности. Универсальный язык может быть либо искусственным, созданным заново, либо естественным, т.е. каким-либо одним из нынешних мировых языков. История разработки универсальных искусственных языков хорошо описана в книге профессора П. Н. Денисова [Денисов 1965]. От Беды Достопочтенного до Ньютона, Декарта и Лейбница создатели искусственных языков мечтали

о всемирном благе посредством исключения войн и конфликтов в обществе с единым языком. Такими же идеями руководствовался и изобретатель эсперанто Лазарь Заменгоф [*Линс* 1999]. Этот язык и является сейчас наиболее распространенным среди искусственных, в некоторых странах есть группы эсперантистов, выпускающих свои печатные издания. Однако руководители информационных служб весьма скептически относятся к перспективе широкого распространения какого-либо искусственного языка как эффективного средства преодоления языковых барьеров. Дело в том, что люди Земли не хотят осознавать себя людьми без Родины, без какого-либо конкретного места, где они родились, воспитывались и живут, где существует общество им подобных. Они не хотят учить язык, который не является родным ни для какого из человеческих сообществ. Иллюстрацией такого рода нежелания и является история эсперанто — как ее обозначил У.Линс, история гонений и преследований.

Второй вариант всемирного языка — один из естественных языков, который возобладает над остальными и станет единым для мирового общения. Надо сказать, что идеи такого рода хорошо известны в интерлингвистике. Дискуссии о едином для человечества языке велись достаточно давно. Было время, когда к мировому господству стремились латинский, французский, немецкий языки. Сейчас первое место по числу публикаций в мире занимает английский, который также преобладает в мировом общении. Однако руководители информационных служб мира также негативно оценивают перспективу его всемирного утверждения. Неслучайно Организация Объединенных Наций имеет несколько языков в качестве официальных. Если принимать единым языком человечества тот, на котором говорит большинство живущих на Земле, то таким языком должен стать китайский или языки народов Индии, на которых говорит наибольшее число жителей планеты. Однако в списке печатных изданий эти языки не занимают ведущих мест. Таким образом, какой-либо один естественный язык не может стать в идеале универсальным средством преодоления языковых барьеров.

Вторым возможным средством преодоления языковых барьеров является изучение иностранных языков. Они составляют обязательный предмет школьного образования во всем мире. В каждой стране свой набор их обязательного изучения, обычно это один или два языка из европейского диапазона. Объединяет школьное преподавание языков во всех странах то обстоятельство, что они изучаются во всех школах мира одинаково плохо, как показывает обследование, проведенное комиссией Европейского сообщества. Безусловно, в этом деле бывают счастливые исключения, и в тех странах, которые находятся на перекрестке мировых коммуникационных и социально-экономических путей, как, например, страны Центральной Европы, где тесно взаимодействуют люди, говорящие на разных языках, школьное языковое образование более эффективно, и коммуникан-

ты владеют несколькими языками в достаточной степени. Но все-таки преодоление языковых барьеров благодаря образованию, которое, между прочим, продолжается и в высшей школе, не носит массового характера и не всегда может обеспечить достаточно эффективный уровень массовой коммуникации.

В свете этого следует признать, что таким средством может оставаться только перевод с одного естественного языка на другой. Данные о распространении переводческой деятельности подтверждают эту точку зрения. Согласно обзору переводческой службы Европейского экономического сообщества объем переводов в мире неуклонно растет, оборот переводческой деятельности, только зарегистрированной, составляет миллиарды долларов в год, и рынок переводов из года в год увеличивается в объеме, не зная «перепроизводства». На арену мирового общения вступают все новые языки.

Перевод и культура

Перевод с одного естественного языка на другой можно рассматривать как процесс создания на другом языке некоторого текста (устного или письменного), эквивалентного по содержанию и способам языкового выражения тексту исходному. Практику и теорию перевода можно отнести к типичным филологическим задачам.

По Ю. В. Рождественскому, исследование текста включает **три ступени**:

– анализ конкретных условий его возникновения, содержащий проблемы авторства и аудитории;

– изучение условий вхождения текста в данную культуру, его роль в данной области культуры, отношения данного текста к другим, проявляющиеся в содержании и языковой форме текста;

– общие исторические закономерности понимания и истолкования текстов на фоне развития культуры, прогресса в знаниях и речевом общении, технического прогресса в создании текстов [*Рождественский* 1979].

Общелингвистические положения теории перевода связаны с общими положениями филологии. Перевод в этом понимании требует углубленного анализа исторической эпохи, культурных ценностей, соотношения реальных, герменевтического анализа текста. Так, сделать древний китайский текст понятным современнику чрезвычайно трудно, тем более трудно добиться того, чтобы этот текст оказал на современника такое же воздействие, как и на своего древнего читателя. Переводчик фактически один такую проблему решить не может. Собственно в его задачу входит, по мнению Ю. В. Рождественского, точная передача стиля. Перевод в этом смысле есть явление социальное, и если в культуре нет соответствия (оно и создается часто благодаря переводу), то нет и собственно перевода как такового, а

требуется истолкование, объяснение, применение других средств воздействия на читателя. Задача перевода состоит в создании текста, построенного на текстовых, а не на языковых эквивалентах. Поэтому идеальным с точки зрения теории перевода в рамках филологической науки является перевод буквальный, подстрочник, такой, в котором есть минимум приспособления культур и свободного толкования. Перевод тем ближе к оригиналу, чем меньше различий в картинах мира двух языков.

В «Переводоведческом словаре» Л. Л. Нелюбин так определяет основные принципы перевода: «1. Перевод должен возможно полно выражать содержание, идею и смысл подлинника. 2. Перевод должен воссоздать художественно-национальное своеобразие оригинала... 6. Правдивый перевод не может не быть буквальным» [Нелюбин 2006]. Мы видим, таким образом, что необходимость, неизбежность и роль буквального перевода признаются с точки зрения как общей филологии, так и сугубо лингвистического подхода к проблемам перевода.

Чрезвычайно важным является также следующее обстоятельство. Осуществляя перевод как истолкование и как стилистическую операцию над текстом, человек-переводчик и редактор вносят неизбежным образом в текст свои собственные знания, понимание и представление о предмете. Поскольку каждое слово языка индивидуально, эти словарные индивидуальности лексических единиц, находящихся к тому же в постоянном развитии и дополнении, соединяются с индивидуальными стилистическими особенностями данного человека. В результате перевод человека всегда индивидуален, практически всегда отклоняется от стиля оригинала вследствие неизбежной индивидуальности переводчика.

Пояснить этот факт можно, рассмотрев труд редактора. Каждый редактор один и тот же текст правит по-своему. В работе «Проблемы машинного перевода» [Марчук 1983] приведен образец машинного перевода, который был передан на постредктирование девяти разным редакторам Всесоюзного центра переводов. Результаты работы показали, что каждый из них вносил индивидуальные изменения в подлежащий редактированию текст. Практически не было двух совпадающих правок. По образному определению Ю. В. Рождественского, такая правка не улучшает, а резко искажает текст, нарушает стиль оригинала. В противопоставление этому чисто машинный перевод нейтрален, он свободен от переводческих предпочтений, от толкований, поскольку машинный интеллект не в состоянии толковать текст, он может только передавать обнаруженные им текстовые соответствия. Именно в этом смысле следует понимать сентенцию о том, что машинный перевод является идеалом всякого перевода по точности и объективности. Объективность машинного перевода объясняется тем, что в составлении словаря и алгоритмов машинного перевода принимает участие достаточно большой коллектив людей,

которые при работе снимают индивидуальные предпочтения стиля и прочие особенности, характеризующие работу индивидуальных переводчиков и редакторов. Более подробно о машинном переводе мы будем говорить в последующих разделах пособия.

Соотношение языков и культур особенно ярко сказывается в многоязычной терминологии. Главным рабочим инструментом переводчика в современной языковой ситуации, когда в преобладающем большинстве случаев он работает с текстами конкретных предметных областей, чаще всего научно-технических, юридических, политических и прочих подъязыков, насыщенными разного рода терминами, которые также изобилуют и в текстах массовой коммуникации, является терминологический словарь. Терминологический словарь для нескольких языков — многоязычный терминологический словарь, варианты которого известны уже с достаточно давних времен, — содержит слова-термины, функционирующие в разных языках и разных культурах. Объем и значение этих слов ясно показывают различия, сходства и несоответствия культурных объектов, обозначениями которых являются данные слова. Можно даже ввести количественные меры расхождения культур, которые (меры) могли бы представить интерес для культурологов и отражаться в количественных распределениях лексики определенных слов и в распределении значений внутри данного лексического массива [*Марчук 1992*].

Место перевода в общей филологической науке, таким образом, характеризуется тем, что он взаимодействует с множеством других наук; в социально-коммуникативном плане и в отношениях с культурой весьма активен и является многоплановым социально- и культурно-обусловленным видом языковой деятельности. Диапазон связей перевода с культурой и филологией чрезвычайно широк.

Перевод и лингвистика

Можно особо отметить роль перевода в лингвистике как науке о языке. Многие лингвистические проблемы тесно связаны с переводом. Так, не требует особых доказательств то обстоятельство, что, например, новые слова являются часто результатом перевода. Однако вопросы моделирования и преподавания перевода требуют изучения и других, более глубоких его связей и отношений с наукой о языке.

Сначала определимся в возможностях моделирования перевода, т. е. описания его в некоторых абстрактных терминах. Рассмотрим вопрос о статике и динамике переводческого процесса.

Целью перевода является передача содержания на уровне языковых средств. Заглубление в смысл необходимо, но только до такой степени, чтобы понять, чем вызвано использование данных языковых средств в оригинале и какие языковые средства должны быть

выбраны для выходного текста. Толкование и перевод текста на выходной язык не должны переходить эти границы.

Языковые средства можно представить себе, с одной стороны, как некоторое упорядоченное собрание, инвентарь слов, морфем, словосочетаний, правил образования выражений языка и их преобразований и правил истолкования. Составляющие такого собрания могут быть перечислены, пронумерованы, записаны в каком-то порядке в хранилище данных. С другой стороны, при распознавании языковых средств выражения в текстах мы должны обращаться в хранилища подобного рода за привлечением конкретных, в данный момент нужных нам сведений — списков слов, правил анализа или синтеза, готовых выражений. В связи с этим можно утверждать, что в моделировании переводческого процесса целесообразно различать **статику и динамику**.

Такое разделение позволяет, во-первых, применить к упорядочению лингвистических сведений методы, разработанные для создания и пользования базами данных, и, во-вторых, ввести понятие алгоритмов работы с этими данными.

В таком подходе, в частности, мы можем определить отношения между единицей перевода и переводным соответствием. Единицу перевода можно определить как основу переводного соответствия. Его, в свою очередь, можно рассматривать как понятие текстовое. Отсюда следует, что процедура его нахождения есть динамика процесса перевода. Единицу перевода, которая базируется на словах или различных комбинациях из слов и синтаксических конструкций, тоже так или иначе привязанных к словам, целесообразно рассматривать как часть системы статики перевода, составляющие которой могут храниться в базах данных.

Следует, однако, отметить, что в таком разделении нет строгих границ. В описании языка декларативные знания не могут быть строго отделены от процедурных в том, что касается процесса перевода. Поэтому можно считать заблуждением представление о том, что перевод осуществляется как алгоритм работы над статическим описанием. Элементы декларативного и процедурного описания языка для целей перевода активно взаимодействуют, что отражается в изложенном выше представлении о статике и динамике перевода.

Многие проблемы переводчик не решает методом простого перебора. Поэтому само по себе разделение данных и алгоритма еще не снимает проблемы оптимального использования накопленных данных. Дело здесь не только в алгоритме, который должен быть оптимальным с точки зрения обращения к базам данных. Сами эти базы должны быть построены таким образом, чтобы можно было оптимизировать алгоритм. Поэтому оптимизация здесь должна быть двух родов: упорядочение данных по отношению к точному типу алгоритма и построение алгоритма, точно отражающего данный процесс при некотором сложившемся расположении данных. Точнее, динамика

определяется статикой, а статика — динамикой. В этом и заключается взаимодействие процедурных и декларативных знаний, без которого моделирование интеллектуальной системы невозможно.

Примерно такой же состав лингвистических и мыслительных действий переводчика выделяют специалисты по синхронному переводу, при котором переводчик как бы мобилизует все умственные резервы для оптимального решения сложной задачи в чрезвычайно сжатые сроки. Поэтому действия синхронного переводчика представляют большой интерес не только для практики перевода, но и для осмысления фундаментальных процессов, происходящих при переводе с одного естественного языка на другой [*Миньяр-Белоручев* 1994, *Виссон* 1999].

Исходя из вышесказанного, можно предположить, что действия переводчика должны моделироваться блоками, в которых должно быть отражено следующее:

- деление процесса, по крайней мере, на две крупные части: обработка составляющих предложения, меньших, чем фраза в целом, и обработка фразы в целом. Фраза есть сложносочиненное или сложноподчиненное предложение, а предложение — простое предложение;

- разделение блоков действия и блоков управления;

- наличие в составе блоков управления страхующих, индикаторных и корректирующих механизмов. Только такие механизмы могут обеспечить переход от предварительного семантического представления к окончательному;

- наличие «банков данных», где хранятся необходимые лингвистические сведения для осуществления процесса анализа и синтеза предложения на выходном языке.

Чрезвычайно важной представляется также обратная связь между результатами анализа частей, меньших, чем предложение, и анализом предложения в целом. Только наличие такой связи дает возможность обеспечить создание цельного перевода.

Блок-схема действий переводчика может быть изображена следующим образом (рис. 1, с. 14).

Правила компоновки предложения есть по существу синтаксические правила выходного языка, по которым следует строить выходное предложение и осуществлять трансформационные преобразования. Трансляционная семантика есть отражение степени понимания переводчиком содержания текста и выражение этого понимания в конкретных действиях по переводу — в подборе лексических эквивалентов и в комбинациях слов в словосочетаниях.

В данной схеме в обобщенном виде отражена связь перевода как действия с лингвистическими сущностями, которые используются при переводе и позволяют создавать перевод уже как объект или предмет, результат действия. Вместе с тем ясно, что эта схема является обобщенным подобием реального состава операций, произво-

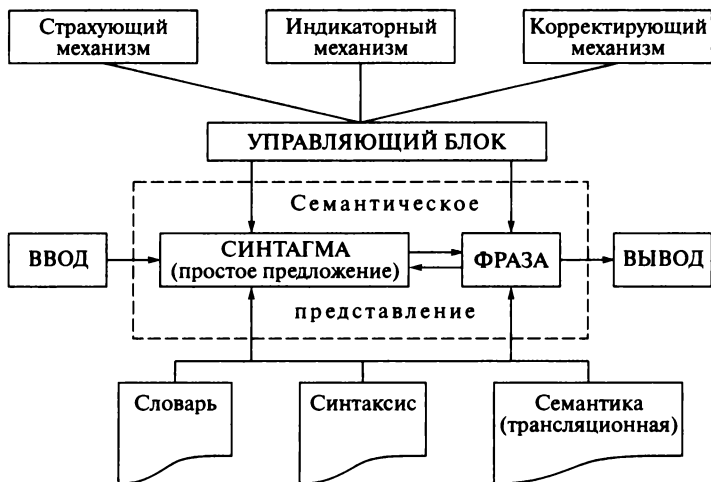


Рис. 1. Блок-схема действий переводчика

димых переводчиком, которые мы не можем непосредственно наблюдать, а можем лишь строить предположения, более или менее правдоподобные, об их составе и содержании. Поэтому то, что изображено на схеме, целесообразно называть «моделью» действий переводчика над исходным текстом.

Понятие модели в лингвистике

Приведенное выше построение является, как нетрудно видеть, моделью перевода как языкового процесса. Понятие модели весьма важно для теории и практики перевода. Моделирование есть способ изучения предмета в тех случаях, когда непосредственное наблюдение его устройства невозможно или крайне затруднительно. Вопросы моделирования перевода с точки зрения формальной теории рассмотрены в книге «Методы моделирования перевода» [Марчук 1985].

Р. Г. Пиотровский вводит два класса моделей — гипотетическую и воспроизводящую [Пиотровский 1979, 1999]. **Гипотетические модели** строятся по следующей схеме:

- формируется гипотеза о внутреннем строении и функционировании оригинала;
- гипотеза оформляется в виде математической формулы или другой абстрактной конструкции, которая выступает в роли теоретического аналога объекта;
- проверяется состоятельность модели по таким характеристикам, как непротиворечивость, исчерпывающий характер описания объ-

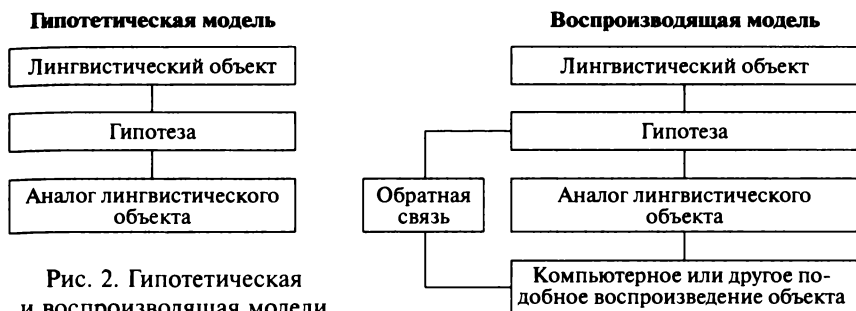


Рис. 2. Гипотетическая и воспроизводящая модели

екта и т. п.; кроме того, определяется, насколько хорошо она объясняет вновь открытые объекты и явления.

Эмпирическая проверка модели включает традиционные лингвистические приемы, такие, как разбор поверхностных структур, данные сопоставительных и сравнительно-исторических исследований и пр. На рис. 2 изображены структуры гипотетической и воспроизводящей моделей.

Примеров **воспроизводящих моделей** можно привести достаточно много. Так, алгоритм анализа морфологической структуры словоформы, действующий самостоятельно или в рамках системы автоматической обработки текста, определяющий структуру и состав элементов, образующих словоформу, может рассматриваться как такая воспроизводящая модель. Система машинного перевода, включающая в себя множество алгоритмов и словарей и осуществляющая перевод с одного естественного языка на другой, также может считаться воспроизводящей моделью, реализующей такую сложную языковую деятельность, присущую человеку, как перевод. Ошибки в переводе, благодаря обратной связи, позволяют совершенствовать алгоритмы анализа и синтеза и пополнять словари системы.

Типология перевода

Прежде чем приступить к моделированию, т. е. к построению конкретных моделей, нужно ориентироваться в вопросе типологии перевода, т. е. решить, какой собственно перевод мы желаем моделировать.

Типология перевода, поскольку она отражает многоаспектность и сложность переводческой деятельности, сама должна быть многоаспектной и сложной. Переводы можно рассматривать с точки зрения:

- формы речевой презентации переводимого текста и текста перевода (письменный перевод письменного текста, устный перевод письменного текста и т. д.);

- временной корреляции процессов порождения оригинального текста и текста перевода (синхронный, последовательный и пр.);
- жанрово-стилистических особенностей переводимого материала (научно-технический, художественный и пр.);
- доминирующего метода и соответствующих ему единиц перевода (модель закономерных соответствий, трансформационная модель и пр.);
- по способу членения оригинального текста на поэтапно переводимые единицы (поморфемный, пословный, абзацно-фразовый и т. п.);
- полноты и качества соответствия текста перевода тексту оригинала (адекватный, неадекватный, реферативный и пр.).

Для моделирования процесса в первую очередь представляется целесообразным разделить на собственно перевод и интерпретацию. Под интерпретацией будем понимать такую часть процесса, в рамках которой переводчик уходит на некоторый уровень, отличающийся от структур чисто языкового выражения. Он пытается передать смысл, содержание, широко используя синонимию средств языкового выражения, знание условий коммуникации, культурные реалии и пр., стремясь в первую очередь добиться эквивалентности в понимании содержания переводимого текста. Переводчик ориентируется в передаче в первую очередь десигнативных и коннотативных значений. В некотором смысле интерпретация есть перевод содержания «своими словами».

Здесь можно обратиться к концепции языкового знака, поскольку переводчик именно им оперирует в процессе перевода. Мы считаем наиболее подходящей для этой цели идею четырехкомпонентной структуры языкового (лингвистического) знака [*Пиотровский 1999*] (рис. 3).

Разные жанры речи (художественный, политический, юридический, научно-технический и пр.) в разной степени могут допускать интерпретацию при переводе. Ясно, что художественный драматический или поэтический тексты гораздо в большей степени могут быть «интерпретированы» при переводе, чем, например, тексты научно-технические или юридические, где отступление от языковых средств должно быть строго контролируемо. В результате классификации текстов по уровню возможной интерпретации и собственно перевода (на уровне чисто языковых средств, без какого-либо толкования) можно построить следующую схему **типологии перевода в аспекте «процесс»** (рис. 4).



Рис. 3. Четырехкомпонентная структура лингвистического знака

Рис. 4. Типология перевода в аспекте «процесс»



В данной схеме видно, что роль перевода на уровне собственно языковых средств повышается в зависимости от типа переводимых текстов.

Типология перевода в аспекте «предмет» важна для моделирования прежде всего затратами индивидуального и коллективного труда. Дело в том, что всякий перевод по своей природе коллективен, поскольку в создании соответствующего текста (письменного) участвуют, по меньшей мере, два человека: переводчик и редактор (корректор). Если в переводе художественного текста роль одного человека — переводчика — преобладающая, а второй участник играет чисто вспомогательную роль, то в таких видах перевода, как научно-технический, коллективный труд более типичен, чем индивидуальный. Так, при переводе технических описаний, например, самолета, которые могут составлять несколько томов, работа разбивается между отдельными переводчиками, после чего требуется согласование результатов их труда. Переводы других жанров носят некоторый промежуточный характер относительно взаимоотношения коллективного и индивидуального вклада. В том, что касается алгоритмов машинного перевода, то они также носят коллективный характер, поскольку составляются группой разработчиков.

Типология перевода в аспекте «предмет» может иметь следующий вид (рис. 5).

Из этой схемы видно, что индивидуальный труд переводчика преобладает в переводе художественном и доля его постепенно снижается в других видах перевода, где все более утверждается коллективный труд.



Рис. 5. Типология перевода в аспекте «предмет»

Машинный перевод в рамках такой типологии безусловно будет признан трудом коллективным, поскольку составление алгоритмов и словарей является делом крупных коллективов разработчиков.

Ю. В. Рождественский несколько иначе строит типологию перевода. От всех других видов речевой деятельности перевод отличается запретом на собственное творчество. Отсюда следует основной критерий его качества. Перевод считается хорошим тогда, когда мысль оригинала полностью доведена до слушателя и читателя в доступной форме.

Это общее требование — полнота мыслей в доступной форме — модифицируется в зависимости от вида словесности, в которой ведется перевод, или от вида перевода, так как каждый вид словесности по-своему формирует мысль в слове и обладает своим критерием понимаемости в аудитории.

Всего определено пять видов перевода.

В переводе документов требуется единственная версия, которая утверждается заказчиком. Исполнители перевода не имеют авторских прав. Существование более чем одной версии перевода помешает действию документа.

При **переводе литературы** допускается множество версий, так как перевод одного и того же оригинала приравнивается к изменяющемуся языковому сознанию читателя литературы. Переводчик пользуется авторским правом, поэтому перевод с точки зрения права не отличим от переделки.

В переводе беседы (в двустороннем устном переводе) переводчик должен добиться согласованности в действиях сторон путем понимания основных мыслей сторон. Перевод предполагает возможность расспроса, переспроса, уточнения намерений и мыслей в диалоге. Языковые формы не столь существенны, но от переводчика требуется поддержание этикета беседы. По ходу беседы он может переформулировать версию перевода.

Перевод синхронный или односторонний (устный) производится как передача монолога, который может иметь письменный прототип. Задача переводчика состоит в точном изложении содержания монолога или, если не позволяют условия, основных положений этого содержания с сохранением основных элементов стиля оригинала.

Информационный перевод текущей научной литературы и документов допускает возможность многих версий, которая редко реализуется. Применение авторского права не обязательно. Широко применяется переделка оригинала, создание рефератов, аннотаций как предварительных версий переводимого оригинала. Перевод может вестись как устно, так и письменно. Деятельность переводчика является частью деятельности информационной службы.

В эти градации укладываются практически все виды перевода текстов (табл. 1).

Таблица 1

Виды перевода	Правила перевода				
	Единичность	Авторское право	Фрагментирование	Оценка по предметной деятельности	Хранимость
Перевод документов	+	—	—	+	+
Перевод литературы	—	+	+	—	+
Двусторонний перевод	—	—	+	—	—
Синхронный перевод	+	—	+	—	—
Информационный перевод	—	+	+	—	+

Рассмотрим, например, перевод научно-технической литературы. Она неоднородна. Такие виды, как научные или технические описания, документация, научные доклады, патенты и т. п., подпадают под категорию перевода документов и поэтому должны существовать в одном варианте, строго соответствовать оригиналу и пр. В то же время научно-технический перевод может также соответствовать категории 5 — информационный, — а также осуществляться в виде беседы, синхронного перевода на научных конференциях и т. п.

Характеристики перевода, вводимые Ю. В. Рождественским, составляют пять непересекающихся признаков. Можно составить таблицу зависимости между видами перевода (характеристиками, которым должны удовлетворять те или иные виды).

Из табл. 1 видно, что по своим характеристикам информационный перевод совпадает с переводом литературы. Однако на него также распространяются правила перевода документов.

Массовая коммуникация — новый этап развития семиозиса

Сначала несколько слов о терминах. Опыт преподавания в вузах дисциплин «Компьютерная лингвистика», «Теория перевода» и других подобных (например, «Автоматическая обработка текстов», «Терминоведение» и т. п.) убедил меня в том, что студенты даже старших курсов не всегда понимают разницу и содержание терминов *семантика*, *семиотика*, *семиозис*. На некотором популярном уровне значение этих слов можно было бы объяснить следующим образом.

Семантика — уровень языка, лингвистическая категория — имеет дело со смыслом, содержанием, понятийным уровнем языковых выражений, в частности и с особенностью лексем. Она выстраивает

ется в ряд с такими категориями, как синтаксис, лексика, морфология, другими чисто лингвистическими понятиями.

Семиотика — наука о знаковых системах, каковыми являются не только естественные или искусственные языки. Это — живопись, музыка, танец, костюм, архитектура, знаки дорожного движения и т. п. Эти системы выражают определенное содержание, воздействуют на умственное восприятие человека.

Семиозис — это то, что происходит в голове у человека, когда он воспринимает тот или иной знак (языковый или любой другой).

Таким образом, современная языковая ситуация и развитие лингвистической науки и смежных наук — семиотики, антропологии, психологии и др. — дают основание ставить вопрос о новом этапе развития человеческого мышления — новом этапе семиозиса.

Продолжим рассмотрение концепции Ю. В. Рождественского. Под массовой информацией понимается совокупное действие массовой периодической прессы, кино, радио, телевидения (прежде всего информационных программ и их окружения). Все эти средства действуют на общей основе информационных агентств, единых принципов финансирования, в рамках определенных принципов информирования, определяемых социальной политикой общества. Так формируется общий смысл массовой информации.

Благодаря совмещению словесных, визуальных и аудиальных средств и постоянству повременного действия массовая информация составляет постоянный *семиотический* (выделено мной. — Ю. М.) фон, на котором разворачивается психическая жизнь людей. Люди ощущают массовую информацию как особую постоянно присутствующую действительность, они приучаются каждый день получать новости, поданные в виде коллажей.

Совокупность сообщений, объединенных коллажем, и структура отдельных сообщений, содержащих факты, их описание, комментариев и художественное сопровождение, не позволяют применить логику к критике как отдельных сообщений, так и коллажей, так как все это оказывает образно-эмоциональное воздействие, к которому логика неприменима. Логической критике можно подвергнуть только направленность сообщений — стратегию информирования. Но выделить и описать эту стратегию можно только специальными аналитическими средствами. Важнейшим средством анализа стратегии является контент-анализ, представляющий собой надежную статистическую оценку содержания текстов.

Ю. В. Рождественский формулирует новые **принципы философии языка XX в.**

Принцип первый. Язык (языковая деятельность) является распорядительной частью семиотической деятельности. Эта деятельность рядоположена системам прогностики, неприкладных искусств (музыка, танец, изобразительное и пластическое искусство), прикладных искусств (архитектура, костюм, дизайн), средствам управления

(меры и измерения, команды и пр.), играм, обрядам, искусственным языкам, в частности языкам науки, техники, математики и пр.

Принцип второй. Сам язык есть сочетание техники создания языковых знаков, составляющих индустрию языка: техника устной речи, материалы и орудия письма, книжная печать, все виды информационных технологий. Каждая технология создания языковых знаков имеет свои потенции для раскрытия смыслов, широты или узости возможностей сообщения между людьми (потенциальный обмен речевых коммуникаций), возможности зафиксировать и упорядочить наличную культуру личности, общества или его частей и организаций. Особенно важна языковая технология для фиксации культуры. Так, устная речь может формировать культуру лишь в памяти людей, письменная — хранить культурозначимые тексты и оперировать ими, создавая их смысловую организацию в виде каталогов и словарей, речь на компьютерах — моделировать все знаковые системы. Компьютерные технологии ставят язык в положение, когда он максимально детально передает глубинные и профессиональные смыслы, свойственные неязыковым знаковым системам.

Третий принцип. Отношения между языком и другими знаковыми системами есть отношения наименования. Слово как лексис становится особенно ответственным, так как правильное именование, лежащее в основе лексических единиц, не только толкует назначение и применение всех вещей, но и определяет их понимание, воспитание людей и управление общественными процессами. Здесь этический центр данной философии языка.

Четвертый принцип. Этические задачи языка — не лгать, не лжесвидетельствовать — не новы, но существенны. В современных условиях возникли новые этические задачи, отвечающие информационному обществу. Это задачи об именах: не создавай имен, в которых лишь частичная правда, не искажай толкованием имен и мыслей другого. Эти новые этические задачи связаны прежде всего с конструированием текстов массовой информации.

Пятый принцип. Для детального учета речи создается новая дисциплина, говорит Ю. В. Рождественский, — общая филология. Общая филология разделяется на три части: учет всех видов речи и их текстовой структуры, представленных в разных культурах; учет того, насколько отдельный человек и все люди в их подразделениях по семейным, профессиональным и другим классам нагружены языком; что такое правильная речь, каковы ее критерии. Учет всех видов речи нужен для их гармонизации между собой.

Шестой принцип. История показывает, что учение о речи — риторика — является показателем стиля жизни. Под риторикой понимается наука о правильном построении речи. Если риторикой пренебрегают, то общество стагнирует.

Седьмой принцип. Философия языка — основание педагогики. От ее характера зависит структура учебного предмета общества и спе-

циального образования, так как языковые действия определяют содержание и методику обучения в их объеме и качестве. Тезаурус образования — самая краткая форма тезауруса культуры.

Восьмой принцип. Эмпирические научные исследования по языку зависят не только от философии языка, но и от организационной формы, когда действуют научно-финансово-промышленные группы, ставящие и решающие конкретные задачи. Пример действия последнего принципа мы видим сегодня в массовой информации, например в телевизионных передачах, которые в решающей степени влияют на современный русский язык, что становится предметом обсуждения в обществе.

Особо следует сказать о роли информатики в массовой коммуникации, о роли новых информационных технологий в современном языковом общении. Эти технологии делают эксплицитными, выносят на поверхность сложные закономерности языковой деятельности человека, которые не всегда могут быть раскрыты на уровне прямого наблюдения, но попытки воспроизвести их для практических целей дают много новой информации об их существовании и функционировании (см., например, [Потанова 2001]).

Глава 2

УНИВЕРСАЛЬНЫЕ ЯЗЫКИ И ИХ РОЛЬ В МАССОВОЙ КОММУНИКАЦИИ

Универсальный язык как средство общения

Проблема универсальных языков достаточно давно привлекала внимание ученых и практиков, поскольку, как можно себе представить из общих соображений, единый язык исключил бы проблему перевода и стал бы общим для всех народов, населяющих Землю. Многие ученые и философы придавали большое значение созданию универсального языка. Наука интерлингвистика специально занимается вопросами искусственных универсальных языков. «Интерлингвистика — раздел языкознания, изучающий *международные языки* как средство межъязыкового общения. Основное внимание обращается на процессы создания и функционирования международных *искусственных языков*, которые исследуются в связи с вопросами *многоязычия*, взаимовлияния языков, образования *интернационализмов* и т. п.» [Лингвистический энциклопедический словарь, 1990].

Одним из примеров продолжающихся работ по созданию искусственных языков является предложенный Д. И. Сотниковым «Язык

международный рациональный арифметический» [Сотников и др. 2004]. В результате синтеза были разработаны так называемые аудиоцифры. Они состоят из двухбуквенных слогов с неповторяющимися буквами: 0 — ро, 1 — ни, 2 — ду, 3 — та, 4 — че, 5 — ся, 6 — бю, 7 — зэ, 8 — вы, 9 — ге. Данная система рассматривается как основа для разработки семейства универсальных мировых языков на базе данной цифровой записи. Авторы идут еще дальше и полагают, что степень сжатия сообщений в системе мировых цифровых языков будет пропорциональна повышению скорости мышления со всеми вытекающими отсюда последствиями. Однако данная идея до сегодняшнего дня большого распространения не получила.

Особое место в мировом языковом творчестве искусственных языков занимает эсперанто [Линс 1999].

Другой способ внедрения и использования одного языка для всеобщего общения — принятие какого-либо естественного языка в качестве универсального. Если исходить из концепции, что таким всеобщим языком должен быть тот, на котором говорит и который является родным для большинства людей на Земле, то таким языком должен стать китайский, на котором говорит большинство живущих в абсолютном выражении. Китайский язык широко распространен во многих странах и районах мира, например в Юго-Восточной Азии. Однако трудности в его освоении заставляют предполагать, что вряд ли он сможет сыграть ведущую роль в решении указанного вопроса. По количеству говорящих на других естественных языках можно, видимо, отметить испанский, русский и английский. На последнем в качестве родного говорит сравнительно небольшая группа людей, однако он по известным причинам глобализации, гегемоном которой выступают США, практически во многих случаях уже играет роль всеобщего мирового языка.

Тем не менее хорошо известно, что многие страны и сообщества не желают всеобщего его распространения, как и глобализации, и предпочитают защищать свой родной язык. Примером в этом отношении может служить Франция, где действуют даже законы, не рекомендуемые употреблять английский там, где есть французский (закон Тубона, например, который запрещает программистам употреблять английский язык вместо французского).

Универсальный язык как фундамент логики общения

Рассмотрим проблему с другой стороны. Математики неоднократно обращались к проблеме универсального языка (см. работы Лейбница и Декарта). Однако они же ставили вопрос таким образом: язык должен строиться по законам логики, он на самом деле существует как отражение логического человеческого мышления и как таковой

поможет не только общению, но и установлению некоторых объективных средств мировой коммуникации.

В. В. Петров и В. Н. Переверзев, математики из Новосибирского отделения Академии наук СССР [Петров и др. 1993], вводят понятие нового статуса языка в связи с развитием новых информационных технологий. Проблемы создания систем машинного перевода, систем обработки и передачи информации решались, как они говорят, стандартными математическими средствами с использованием элементарной логики и простейших лингвистических сведений о языке.

При этом математики отмечают, что нет должной ясности в общих вопросах, касающихся самой сущности языка. В течение всех предыдущих эпох язык был в большинстве случаев предметом отвлеченных рассуждений, а теперь требуются точные и однозначные ответы на поставленные практикой и теорией вопросы. Это особенно важно на уровне проблем машинной формализации естественного языка.

В настоящее время нет достаточно мощной формальной теории естественного языка, в рамках которой можно было бы дать адекватные ответы на поставленные вопросы. Такая теория еще только разрабатывается, говорят упомянутые математики. Она предполагает ряд новых исследований на стыках разных наук — лингвистики, психологии, искусственного интеллекта. При этом пока остается открытым вопрос: как строить такую теорию — на основе логики предикатов или на какой-либо другой?

Универсальный язык как основа автоматизации перевода

Может ли искусственный универсальный язык в любой его форме стать основой автоматизации перевода?

С начала исследований по машинному переводу такой вопрос ставился не только в практической форме (работы по проекту ЕВРО-ТРА), но и главным образом с точки зрения теории. Однако в части автоматизации и построения систем машинного перевода были предприняты попытки использовать некоторые универсальные языки смысла для применения их в качестве основы для перевода на многие естественные языки. Это не привело к созданию практических систем автоматизированного или автоматического перевода. Хотя исследования общелингвистических закономерностей, которые были необходимы при попытках построения таких систем в русле этого теоретического направления, безусловно служили определенным вкладом в развитие интерлингвистики — науки об универсальных закономерностях человеческих языков.

Трудность формализации семантики языка до такой степени, чтобы она могла служить основой построения алгоритмов перевода, заключается в том, что между смыслом и содержанием текста есть без-

условная разница. А. И. Новиков описывает результаты проведенного им эксперимента, когда группе из десяти человек (студентов) были даны три текста, объемом около страницы каждый, с заданием указать на то, в каких из них смысл соответствует содержанию: описание походов Александра Македонского (исторический текст); текст, описывающий некую химическую реакцию при получении какого-то вещества; художественный текст, описывающий беседу двух молодых людей — юноши и девушки — во время поездки в поезде: они смотрят в окно и обмениваются впечатлениями о пейзаже и том, что видят из окна поезда. На вопрос о том, соответствует ли смысл содержанию первого текста, исторического, большинством респондентов был дан ответ, что смысл соответствует содержанию примерно наполовину. Содержание — описание конкретных походов Александра Македонского; смысл — он был великим завоевателем, однако порой был жесток. Во втором тексте, химического характера, смысл был определен как полностью соответствующий содержанию. Третий текст, художественный, квалифицирован как представляющий полное несоответствие смысла и содержания. Содержание — обмен замечаниями по поводу того, что видно из окна поезда; смысл — выяснение отношений между молодыми людьми: «Почему ты меня разлюбил (разлюбила)?» и т. п. [Новиков 2002]. При этом никаких предварительных определений того, что является смыслом, а что — содержанием, не давалось, эти категории определялись информантами интуитивно.

На расхождение между смыслом и содержанием обращал внимание В. А. Звегинцев, когда утверждал, что смысл предложений *Охотник выскочил из-за куста и ударом ноги убил волка*, *Волк был убит ударом ноги выскочившего из-за куста охотника*, *Из-за куста выскочил охотник и ударом ноги убил волка* и других возможных парафраз этого предложения отнюдь не один и тот же, а зависит от общего контекста, в который входят эти предложения [Звегинцев 1976].

В связи с этим использование смысла (содержания) как основы для построения алгоритмов машинного перевода связано со значительными трудностями, и пока еще мы не имеем ни одной действующей системы машинного перевода, которая была бы построена на принципиальном использовании формализованного смысла.

Глава 3

ТЕРМИНОЛОГИЯ, ТЕРМИНОВЕДЕНИЕ, ТЕРМИНОГРАФИЯ

Мы выносим проблемы терминологии и терминологической лексики в заголовок и начало данной главы, посвященной моделированию перевода лексики, по причинам большой и возрастающей роли

терминологии в современном языковом общении. Напомним, что в концепции Ю. В. Рождественского современная массовая коммуникация состоит из трех составляющих: массовой информации, информатики (научно-техническая информации) и современной рекламы. Кроме того, в обоснование постановки на первое место проблемы терминологии можно сослаться на современные исследования, по которым в антропологистике первое место занимает именно терминология [Гринев-Гриневиц и др. 2005]. Это и понятно, можно обоснованно предположить, что первыми словами, которые произнес первобытный человек, были термины — *дубина, огонь, копье, очаг, пища* и пр., и лишь потом словарный состав расширился. В современном языкознании уже преодолена точка зрения на то, что термин не относится к общепринятому литературному языку, а стоит как бы на периферии словарного состава. Рассмотрим лексический уровень современного естественного языка.

Лексический уровень системы языка

П. Н. Денисов, анализируя русский язык, высказывает кардинальные соображения относительно лексической системы любого языка. По его определению, язык представляет собой длительно эволюционирующую систему систем, не только потому, что по мере развития и усложнения обслуживаемой им культуры растет, ширится, разветвляется и дифференцируется лексическая система, но и потому, что она эволюционирует вместе с развитием грамматической и фонетической систем. Развитие основных «уровневых» систем языка — фонетической, грамматической и лексической — при их безусловной взаимосвязи все-таки совершается с известной долей самостоятельности хотя бы в смысле скорости изменения, степени подвижности и характера восприятия внешних влияний со стороны «среды», т. е. общества, народа, его духовной и материальной культуры. Для системы, по мнению П. Н. Денисова, характерно следующее: она представляет собой целостный комплекс взаимосвязанных элементов; образует особое единство со средой; как правило, любая система представляет собой элемент системы более высокого порядка; элементы любой системы, в свою очередь, выступают как системы более низкого порядка.

В качестве примера устойчивости и в то же время изменчивости лексической системы можно привести группу слов, обозначающих разные роли человека в семье. Такие слова, как *мать, брат, сестра, сын* и прочие, имеют древнее, праиндоевропейское происхождение и прекрасно существуют в наше время, в то время как другие слова этой же группы, например *золовка, деверь, шурин* и т. п., вряд ли понятны современному человеку. В более широком смысле частные системы можно признать элементами глобальной системы; в более диф-



Рис. 6. Лексический уровень системы языка [Денисов 1974]

ференцированном рассмотрении элемент глобальной системы можно исследовать как систему или, по крайней мере, микросистему. Основные качественные изменения лексической системы современного русского литературного языка вызываются тремя причинами: бурным развитием терминологии в век научно-технической революции; вхождением в словари новых слов нетерминологического характера; дальнейшим ростом стилистически маркированной лексики, в особенности самого многочисленного ее разряда — разговорных слов (рис. 6).

Что касается чтения специальных текстов, то здесь следует уточнить точку зрения П. Н. Денисова, поскольку обилие терминов в специальных текстах может затруднить их понимание. В этой связи можно сослаться на мнение С. В. Гринева-Гриневица [Гринева-Гриневица 2008]. Он пишет, что, по данным науковедения, в среднем каждые 25 лет число научных дисциплин удваивается (по данным Британской энциклопедии, в XX веке появилось свыше 2000 новых научных дисциплин). Это еще раз подчеркивает важность третьего принципа новой философии языка, сформулированного Ю. В. Рождественским, а именно — правильное именование определяет правильное понимание и использование предмета.

Термины и терминология в лексике языка

Уточним содержание терминов «терминология», «терминоведение», «терминография». Слово **«терминология»** многозначно. Оно имеет три значения: наука о терминах, область прикладной (можно теперь сказать «и теоретической») лингвистики; система терминов конкретной предметной области (например, медицины); система терминов конкретного научного направления в рамках предметной области (например, термин «фонема» в понимании ленинградской и московской научных направлений языкознания). Слово **«терминоведение»** однозначно определяет науку о терминах, здесь нет других

значений. Первое значение слова «терминология» как науки о терминах в настоящее время все чаще заменяется понятием «терминоведение». Это связано с тем, что языкознание признает термин полноправной лексической единицей всякого естественного языка и больше не считает термины какой-то языковой периферией. Значение слова «**терминография**» также определено: область лексикографии, которая занимается составлением терминологических словарей. При этом понятие «терминоведение» трактуется разными специалистами с различных точек зрения.

С. В. Гринев [Гринев-Гриневиц 2008] полагает, что основная цель терминоведения — изучение особенностей и закономерностей образования и развития терминологий для выработки рекомендаций по их совершенствованию и наиболее эффективному использованию. Для достижения этой цели, по его мнению, необходимо решение ряда теоретических и практических задач. К **теоретическим задачам** относятся:

- установление и описание основных типов специальных лексических единиц;
- выработка общих методов описания и анализа терминологий;
- определение общих свойств терминов и терминологий;
- исследование основных типов называемых терминами понятий и связей между ними, семантических особенностей и проблем терминов и терминологий;
- изучение структурного и словообразовательного состава терминов и терминологий;
- исследование особенностей зарождения, образования и развития терминологий различных областей знания;
- анализ особенностей функционирования терминов и терминологий в специальной речи и в современных автоматизированных системах;
- определение роли терминов и терминологий в познании, мышлении и развитии знаний, а также в подготовке специалистов;
- совершенствование теоретических основ создания различных типов словарей специальной лексики.

Если проанализировать приведенные выше задачи терминоведения, то можно выделить в приведенном списке занятия для трех видов специалистов: лингвистов, философов и специалистов-предметников, т. е. специалистов конкретных областей знания.

С. В. Гринев-Гриневиц далее выделяет **практические задачи** терминоведения, к которым относит следующие:

- разработка методики нормализации (рекомендации и стандартизации) и создания терминологий в разных областях знаний;
- установление критериев и принципов отбора и обработки специальной лексики при создании различных специальных словарей;
- разработка методов терминологической работы при создании информационно-лингвистического обеспечения современных ком-

пьютеризированных систем — АСУ, баз и банков данных, баз знаний и интеллектуальных систем (сюда следовало бы также добавить терминологические проблемы машинного перевода как особого сложного вида применения новых информационных технологий к решению лингвистических задач);

- разработка методов терминологического редактирования и терминологической экспертизы;

- разработка методов, приемов и рекомендаций по переводу терминов.

Точные границы терминоведения, основные сведения по структуре, аспектам и отраслям современной науки о термине, информация о вкладе отдельных ученых-терминологов, адекватная картина понятийно-терминологического аппарата, неразрешенные проблемы терминоведения — все это дано в энциклопедическом словаре «Общее терминоведение» [см. *Татаринов 2006*].

Терминография

Решением практических задач терминоведения занимается терминография — его составная часть, занимающаяся созданием терминологических словарей.

Терминография является частью лексикографии [*Марчук 1992*]. Термин «лексикография» имеет три основных значения: наука о составлении словарей; составление словарей как описание лексики того или иного языка; совокупность словарей, относящихся к данному языку или к области знания. Иногда этот термин понимается как совокупность словарей, составленных в данной стране.

Как система знаний, лексикография включает в себя объект и методы исследования. Объектом исследования лексикографии являются словари, составляемые для пользования человеком или для использования в компьютерной обработке текстов.

Словари разнообразны по своим типам, зависящим от их задач и целевого назначения. Они выполняют различные функции, среди которых можно выделить следующие:

- регистрация в словесной форме объективных данных о внешнем мире в свойственной данному языку и эпохе форме восприятия мира (энциклопедические и толковые словари);

- упорядочение в понятийно-словесной форме субстанции содержания данного языка (идеологические словари, словари-тезаурусы);

- нормализация словоупотребления с целью облегчения языковой коммуникации (нормативные и терминологические словари);

- перевод с одного языка на другой или на другие языки (двухязычные и многоязычные переводные словари);

- вспомогательные операции для пользования языком (словари-справочники, номенклатуры и пр.) [*Марчук 1976*].

Типология словарей

Рассмотрим кратко основные виды словарей, обращая особое внимание на следующие три аспекта:

- цели и задачи словаря;
- тексты, которые охватывает словарь, или подъязык, который описывается данным словарем;
- категория пользователей.

Три эти аспекта могут служить основой для их типологии. Л. В. Щерба выдвигал следующие основные положения для создания общей типологии словарей [цит. по *Денисов 1974*]: а) словарь академического типа — словарь-справочник; б) энциклопедический словарь — общий словарь; в) тезаурус — обычный толковый или переводной словарь; г) толковый или переводной словарь — идеологический словарь; д) толковый словарь — переводной словарь; е) исторический — неисторический словарь.

Рассмотрим вопрос отбора лексики в словари. Исходный список слов — словник — определяется типом словаря и в значительной мере интуицией и компетентностью лексикографа, что дает основание рассматривать лексикографию не только как науку, но и как искусство. Источниками, из которых отбирается словник, являются:

- уже существующие словари такого же типа, как и проектируемый, или сходных типов;
- письменные тексты данной области знания;
- собственные наблюдения авторов над речью и их собственный языковой опыт.

На отбор лексики существенно влияют норма, а также объем словаря и его целевое назначение.

Рассмотрим влияние нормы. Обычно на ее основе принимается решение о том, включать ли данное слово в словник или нет, а если включать, то какими пометами его сопровождать. Основными критериями нормы считаются:

- употребление данных языковых фактов писателями прошлого и длительность существования этих фактов;
- соответствие языковых фактов закономерностям и тенденциям языка;
- распространенность языкового явления [*Берков 1973*].

Однако не всегда эти критерии соблюдаются. Так, как отмечает В. П. Берков, употребление писателями отдельных слов может содержать прямые ошибки. Реализации второго критерия — соответствие языковых фактов внутренним закономерностям языка — препятствуют следующие факторы: возможность различного оформления аналогичных образований; отсутствие точного определения продуктивных моделей; трудность выявления внутренних законов языка применительно к лексике. Распространенность какого-то лингвистического факта, употребление языкового явления, также не всегда может

служить показателем его нормативности. В связи с этим можно отметить, например, сленг.

Ю. С. Маслов рассматривает следующие типы лингвистических словарей.

Толковый словарь дает толкование значений слов (и устойчивых словосочетаний) какого-либо языка с помощью слов этого же языка. Толкование дается посредством логического определения концептуального значения (например, *накалиться — нагреться до очень высокой температуры*), посредством подбора синонимов (*назойливый — надоедливый*) или же в форме указания на грамматическое отношение к другому слову (*прикрывание — действие по значению глаголов прикрывать и прикрываться*).

Толковым словарям противопоставляются переводные словари, двуязычные и многоязычные. В них вместо толкований даются переводы значений слов на другие языки.

К общим словарям относят также и те, которые рассматривают все пласты лексики, но под каким-либо специфическим углом зрения. Здесь выделяются словари: частотные, показывающие частоту употребления данного слова; грамматические, дающие подробную грамматическую характеристику слов; словообразовательные, указывающие членение слов на составляющие их формологические элементы, словари сочетаемости и др.

Этимологические словари содержат сведения о происхождении и первоначальной мотивировке слова.

Терминологические словари в данной классификации образуют особую группу. Будучи собраниями слов естественного языка, они отражают основные признаки, которые служат основанием для классификации нетерминологических словарей. В то же время они особым образом относятся к концептам, лежащим в основе терминологий. Они организуют эти терминологии и одновременно отражают их в языке. Функционирование терминов в речи и текстах связано с характером современной языковой ситуации в мире и с особенностями современного языкового общения.

Современная языковая ситуация.

Проблемы деловой прозы

Языковая ситуация в современном мире характеризуется, с одной стороны, новым этапом семиозиса, как указано выше, с другой стороны, преобладанием так называемой деловой прозы. Если исходить из тройки, составляющей современную массовую коммуникацию, а именно — массовая информация + информатика + новая реклама, то современная языковая ситуация характеризует в основном две составляющие — массовую информацию и информатику, понимаемую как научно-техническое информирование.

Два следующих фактора характеризуют современную языковую ситуацию:

- увеличение числа национальных языков, входящих в сферу мирового общения;
- преобладание деловых документов над ранее господствовавшими текстами религиозного содержания, художественной литературы, мемуарами и т. п.

Рассмотрим эти два фактора более подробно.

По состоянию на сегодняшний день в мире насчитывается не менее 3 000 языков, на каждом из которых говорит от миллиарда до 50 000 человек. Наиболее распространены (по числу говорящих на них от рождения) языки китайский, индийские (хинди и бенгали), испанский, португальский, русский. Однако литературный и «печатный» выход каждого из языков не находится в пропорции с количеством людей, говорящих на данном языке как на родном. Так, китайский язык, самый распространенный на Земле по числу говорящих на нем людей, охватывает только 1 % мировой печатной продукции. По данным ООН, в мире ежегодно выходит около 500 000 наименований печатной продукции. По языкам эти издания распределяются следующим образом: 120 000 — на английском языке; 60 000 — на русском; 50 000 — на французском; 40 000 — на немецком и т. д. Примерно по 10 000 изданий выходит на корейском, польском, датском, норвежском, итальянском и других языках [по данным Better Translation for Better Communication 1983]. Несмотря на то что эти данные двадцатисемилетней давности, нет оснований предполагать, что основное соотношение каким-либо существенным образом изменилось. Можно только подвергнуть сомнению второе место русского языка в связи с распадом СССР и появлением многих новых национальных языков в границах бывшего Советского Союза.

В связи с таким обилием национальных языков и растущим участием их в международной коммуникации преодоление языковых барьеров становится все более насущной необходимостью. Как уже было показано выше, преодоление языковых барьеров наиболее эффективно осуществляется именно переводом.

Что переводится в мире? По данным ООН, по типам переводимых текстов переводы характеризуются следующим образом:

коммерческие	35,4 %	аудиовизуальные	2,1 %
промышленные	21,0 %	учебные	1,5 %
научные	20,0 %	литературные	0,3 %
юридические	9,3 %	прочие виды	6,9 %
пресса и текущая информация	3,5 %	Итого:	100 %

Из этих данных видно, что объем литературного перевода незначителен по сравнению с переводом коммерческим, научным и промышленным, общий объем которых почти доходит до 80 %.

Какие из этих видов перевода можно отнести к деловой прозе? Ответ достаточно ясен. Первые четыре, а именно переводы коммерческие, промышленные (научно-технические), научные, юридические, составляющие в общем объеме 85,7 %. Пресса и текущая информация могут содержать как тексты деловой прозы, так и не относящиеся к таковой, например фельетоны, анекдоты, пародии, стихи и т. п. Однако в общем объеме эти тексты составляют всего 3,5 %. Аудиовизуальный перевод — так квалифицирован, по-видимому, синхронный перевод, перевод беседы, последовательный устный перевод и другие подобные виды — может содержать тексты деловой прозы в каком-то объеме, но этот вид перевода также в общем имеет небольшой процент содержания. Учебный перевод вряд ли стоит относить к деловой прозе, равно как и прочие виды перевода.

Таким образом, мы видим, что деловая проза — в широком понимании этого термина — занимает основное место в общем объеме переводов в мире. Именно поэтому мы и посвятили данный раздел терминоведению, терминологии, терминографии, поскольку деловая проза в широком понимании не может обойтись без этого типа лексических единиц.

Терминологические словари

По данным Всесоюзного центра переводов научно-технической литературы и документации, который провел специальное исследование терминологической лексикографии, в мире каждый день публикуется один специальный словарь [Татаринов 2006].

Терминологический словарь является необходимым инструментом для осуществления научного и научно-технического перевода. В мире нет переводчиков научно-технической литературы и документации, которые могли бы обойтись без этих словарей. В современных текстах много новых научно-технических терминов, отражающих развитие науки и техники, сопровождающее мировую цивилизацию и прогресс во всех областях жизни. Например, в современном русском языке появились такие термины, как *менеджер*, *мерчендайзер*, *транспарентность*, *волатильность*, *позиционировать*, *девелопер* и многие другие. Некоторые из них безусловно являются терминами, например, банковского дела: *маржа*, *волатильность*, *ликвидность*. Другие же слова, такие, как *позиционировать*, *транспарентность* и пр., приобрели статус общеупотребительных, хотя встречаются в основном в политических текстах и могут быть не понятны широкому читателю и слушателю. Многие из них еще не вошли в обычные словари.

В целом лексическое богатство каждого естественного языка увеличивается из века в век. Некоторая иллюстрация этого положения: первый словарь английского языка, опубликованный как результат научного исследования языка в 1604 г., содержал 3 000 слов. Словарь

Джонсона 1750 года — 43 500 слов. В 1973 году Краткий Оксфордский словарь — 163 000 слов. Во Франции Универсальный словарь Фуретье 1690 года — 50 000 слов. Словарь французской Академии 1964 года — 25 000 только наиболее высокочастотных слов. Энциклопедия Гран Лярусс в десяти томах содержит все слова французского языка — 450 000 [Марчук 1992].

В работе «Основы терминографии» [Марчук 1992] приводятся данные Всесоюзного центра переводов научно-технической литературы и документации о количестве терминологических словарей в мире и распределения этих словарей по отдельным областям науки и техники в разных странах. Каждая страна имеет свой набор терминологических словарей, который в существенной степени характеризует промышленность и направление развития ее хозяйства. Так, во Франции, при наличии большого количества терминологических словарей по разным направлениям науки и техники, большое внимание уделяется словарям медицинским. В Китае преобладают словари по биологии. Индия характеризуется обилием двуязычных терминологических словарей для перевода. Промышленному развитию Японии предшествовал бум в публикации терминологических словарей разных областей промышленности и науки.

Какое значение эти данные имеют для моделирования перевода? Нет сомнений в том, что при построении систем и алгоритмов перевода для конкретных языковых пар безусловно следует учитывать наличие, объем и значение терминологической лексики соответствующих предметных областей.

Глава 4

ТЕРМИНОЛОГИЯ И ОБЩЕУПОТРЕБИТЕЛЬНАЯ ЛЕКСИКА В ПЕРЕВОДЕ

Общеупотребительная и общенаучная лексика

В научно-техническом переводе наибольшие трудности возникают при передаче лексических единиц оригинала. Как показывает анализ англо-русских технических переводов, основной процент переводческих ошибок встречается в сфере лексики — наиболее динамичном в семантическом отношении уровне языка науки и техники. Здесь часты как неправильный перевод терминов, так и случаи неверной передачи значений общенаучных слов, причем последние преобладают.

Понятие «значение» трактуется по-разному. Например: *заглушка* и *затычка*. Можно сказать, что здесь одно лексическое значение. Но: *каждой бочке затычка*. В то же время если слова используют-

ся как термины, то совершенно разные. Другой пример: *стол* как меню и как предмет мебели. Вроде непохоже, что одно значение. Но: есть общая сема: *сидеть за столом* значит *есть*. Так что понятие «значение» весьма неопределенно в традиционной лингвистике, но в формальной, компьютерной, автоматической — разные слова — разные значения.

Проф. Л. И. Борисова специально исследовала вопрос о соотношении общеупотребительной лексики и терминологии в переводах научно-технических текстов на материале большого количества переводов [Борисова 2005]. Рассмотрим ее выводы и результаты исследования, проведенного на материале англо-русских и русско-английских научно-технических переводов.

В указанной работе Л. И. Борисовой приведены результаты изучения общенаучных слов, представляющих затруднения при переводе научно-технических текстов, таких, как *candidate*, *history*, *feature*, *beauty*, *handle* и т. п. Общенаучный слой лексики имеет большое значение для смысловой и структурной организации текста. Общенаучные слова являются средством обозначения интердисциплинарных понятий, а также связующим звеном лексического состава научно-технических текстов. Они регулярно, но с разной частотностью, встречаются в различных подъязыках, причем довольно часто в специфических значениях.

Из всей совокупности работ по стратификации лексического состава научно-технических текстов можно выделить два основных подхода: классификацию лексики на три крупных слоя — термины, общенаучные слова, общеупотребительные слова; классификацию лексики на два крупных слоя — термины и слова-нетермины (которые разными исследователями называются по-разному, например: общенаучные слова, общая научно-техническая лексика, нетерминологическая лексика и т. п.). Однако трудно найти различия между ними. Например, прилагательное *available* встречается во всех трех упомянутых списках, составленных разными исследователями.

Л. И. Борисова придерживается второго подхода, поскольку четкие критерии стратификации нетерминологической лексики не выработаны, и к общенаучной лексике она относит все слова-нетермины, функционирующие в языке науки и техники и имеющие определенную семантическую специфику.

В соответствии с данным подходом за общенаучные слова Л. И. Борисовой принимаются любые слова-нетермины, функционирующие в языке науки и техники, обладающие определенной качественной спецификой по сравнению со своими характеристиками в общем языке, значения которых в основном не меняются от подъязыка к подъязыку, обозначающие как основные общенаучные понятия, так и самые разнообразные и различные понятия общенаучного обихода, а также выполняющие связующую и организующую функции в языке науки и техники.

Всю совокупность общенаучных лексических средств, функционирующих в языке науки и техники, можно условно разбить на три основных подслоя, исходя из их информационной значимости, а также функции, выполняемой ими в научно-технических текстах.

Первый подслоя можно обозначить как ядерную часть (центр, ядро, каркас, основу) общенаучной лексики. Его составляют общенаучные слова, обозначающие основные общенаучные междисциплинарные понятия. Это — наиболее существенная часть общенаучного словаря. Сюда относятся полноточные слова, непосредственно описывающие объекты научного и технического исследования и разработок. Они составляют основу слоя общенаучной лексики, например: *application, development, background, analysis, arrangement, history, generate, concept, technique, equipment, philosophy, extension* и т. п.

Второй подслоя — слова общенаучного обихода. К нему относятся слова, обозначающие самые различные понятия, имеющие отношение к сфере деятельности человека в науке и технике. Многие из них либо являются синонимами ядерных общенаучных слов, либо используются в аналогичной функции для более детальной характеристики предметов научно-технической деятельности человека, например: *advent, arena, attraction, avenue, beauty, birth, bottleneck, bread-and-butter, exotic, impressive* и т. п. Подобные слова являются неотъемлемой частью лексического состава научно-технических текстов, участвуют в построении научного изложения, но не определяют его в такой степени, как ядерная лексика. Этот подслоя является наиболее открытым. Он представляет собой как бы стилистическую вариацию. Вариативность в текстах слов этого подслоя значительно выше, чем ядерной лексики. Слова варьируются в зависимости от тематики.

Наконец, **третий подслоя** — базовые строевые (грамматические) общенаучные слова, являющиеся связующим и организующим элементом языка науки и техники. К этому подслою относятся наиболее распространенные и часто употребляемые слова, образующие лексический фундамент языка в целом, не варьирующие в различных функциональных стилях. Этот подслоя представляют в основном служебные лексические элементы: артикли, предлоги, союзы, вспомогательные глаголы, местоимения и пр.: *a, the, any, or, do, be, once* и т. п. Исследование этих слов в научно-технических текстах показало, что они не столь уж нейтральны и приобретают определенную специфику в процессе функционирования в языке науки и техники, что требует дальнейшего изучения.

Наряду с этими можно выделить и другие группы слов, например пласт общенаучной интернациональной лексики по определенной общности характеристик в разных языках — семантической, графической, фонетической и морфологической.

В общем и целом можно отметить, что состав общенаучной лексики неоднороден. Он включает как регулярные и устойчивые, так

и менее регулярные и окказиональные лексические средства, обладающие различной информационной значимостью, которые образуют систему общенаучной лексики, семантически и стилистически гибкую и динамичную, являющуюся интердисциплинарным лексическим фундаментом научно-технических текстов, обслуживающую любые подъязыки науки и техники, всю сферу общенаучной коммуникации.

Переводческая классификация общенаучных слов

Продолжим рассмотрение результатов Л. И. Борисовой. В процессе отбора лексики проводилось сопоставление контекстной семантики общенаучных слов с информацией словарей, которое показало, что значения, актуализирующиеся в научно-технических текстах, не представлены в общих двуязычных переводных словарях и во многих случаях не зафиксированы толковыми словарями английского языка. При проверке общенаучной лексики по Большому англо-русскому словарю (БАРСу) было установлено, что некоторые из отобранных слов в нем вообще отсутствуют, а общенаучные эквиваленты многих слов резко расходятся с эквивалентами БАРСа.

В результате анализа всей отобранной общенаучной лексики были выделены пять групп слов, вызывающих затруднения при переводе научно-технических текстов: общенаучные слова, отсутствующие в БАРСе; «ложные друзья переводчика»; слова широкой семантики; низкочастотные слова с узкой семантикой; стилистически окрашенные лексические средства.

Прежде всего обращают на себя внимание **общенаучные слова**, функционирующие в научно-технических текстах, но отсутствующие в большинстве общих переводных словарей, такие, как *configure*, *overpredict*, *underpredict*, *emerging*, *demonstrated* и т. п. Большинство из подобных неологизмов создаются по традиционным моделям словообразования английского языка, включают известные корни, аффиксы, суффиксы и префиксы. Однако при их переводе на другой язык возникают затруднения, при этом значения однокоренных слов, имеющих в словаре, часто не помогают. Некоторые из таких неологизмов имеют очень высокую частотность. Например *specifically* и *configure*; *specifically* обычно встречается в начале предложения и выполняет функцию логической связки; на русский язык передается широким спектром переводных эквивалентов, к которым относятся: *в частности, например, так, характерно, что, точнее говоря, более конкретно, более определенно, дело в том, что, в связи с этим, с целью* и пр. Наиболее частым является эквивалент *в частности*.

Ко второй группе общенаучных слов относятся «**ложные друзья переводчика**». Часто в этой роли выступают слова-интернациона-

лизмы, употребляемость которых является весьма высокой. Перевод их достаточно закономерен в научно-технических текстах. У них практически никогда не реализуются их основные интернациональные значения. Это такие слова, как *candidate, elegant, history, dramatically, menu, provocative, talent* и т. п. В научно-технических текстах требуется выбор иных переводных эквивалентов.

К третьей группе общенаучных слов относятся **слова с очень широкой семантикой** типа *handle, feature, associate, meaningful, available, support, treat, predict* и т. п. Они имеют как довольно устойчивые переводные эквиваленты, так и очень широкий спектр самых различных нерегулярных эквивалентов. Например, у глагола *handle* регистрируется 30 новых регулярных устойчивых переводных эквивалентов, отсутствующих в БАРСе, у глагола *feature* — 15, у прилагательного *meaningful* — 35 и т. д. Так, последнее часто выступает в значении *надежный* — *meaningful measurement* и т. п.

Четвертая группа — **слова с небольшим объемом значений**, характеризующиеся спецификой, не отраженной в обычных словарях, например: *advantageous, jeopardize, desired* и т. п. Они выступают в научно-технических текстах в определенных конкретных значениях и имеют ограниченное число переводных эквивалентов. *Outgrowth* переводится как *развитие, появление, продолжение*. Общеупотребительные значения у этих слов практически не реализуются.

К пятой группе относятся **стилистически окрашенные лексические средства**. Это слова субъективной оценки, которые также требуют соответствующих переводных эквивалентов, например: *excellent, remarkable, extremely* и т. п. При их переводе надо ориентироваться на стилистические нормы русского языка.

В связи с передачей контекстных значений английских общенаучных слов целесообразно рассмотреть, в какой степени русские переводные эквиваленты передают значения иностранных слов, какие компоненты значений воспроизводятся в научно-техническом переводе.

Контрольные вопросы

1. Какое влияние оказывает перевод на культуру общества?
2. В чем заключается новый этап семиозиса?
3. В какой мере способствуют универсальные языки преодолению языковых барьеров?
4. Каково содержание терминов «терминология», «терминоведение», «терминография»?
5. Какие типы словарей вы знаете?
6. Какие вы можете выделить особенности перевода общенаучных слов в научно-техническом переводе с английского языка на русский?
7. Как влияют заимствования (и влияют ли) на развитие лексической системы языка?

РАЗДЕЛ II

СТАНОВЛЕНИЕ И РАЗВИТИЕ МАШИННОГО ПЕРЕВОДА

Глава 5

ПОЯВЛЕНИЕ ИДЕИ МАШИННОГО (АВТОМАТИЧЕСКОГО) ПЕРЕВОДА

Как начинался машинный перевод

Возможность механизации такого казавшегося рутинным умственного труда, как перевод, давно привлекала внимание ученых. В 1933 г. в СССР было запатентовано механизированное переводное устройство П. П. Смирнова-Троянского [*Панов и др.* 1956]. Однако на машинный перевод это устройство не оказало никакого влияния, поскольку в нем было реализовано слишком примитивное представление о сущности переводческого процесса. Собственно машинный перевод начинается с применением электронных вычислительных машин (ЭВМ), как раньше назывались нынешние компьютеры, для обработки текста на естественных языках (см. также [*Марчук* 1983]). Эта идея основывалась на успехах, достигнутых в применении ЭВМ к криптографическим задачам. В дискуссии между У. Уивером и А. Д. Бутом в 1946 г. было высказано предположение, что «основные элементы языка могут быть обнаружены с помощью устройств, разработанных во время Второй мировой войны и предназначенных для расшифровки секретных кодов противника» [*Бут и др.* 1957]; (см. также [*Early Years in Machine Translation 2000*]). После публикации работ К. Шеннона по математической теории связи, оказавших большое влияние на развитие математической теории информации, были сделаны попытки применить к анализу языка инструментарий средств, пригодных для обнаружения общих свойств кодов, и машинный перевод рассматривался как еще одна область применения техники дешифрования. При этом важно, что предполагалось осуществить обнаружение на первых порах самых общих свойств кода. В меморандуме У. Уивера — одном из исходных документов, с которых началось развитие машинного перевода, — приводится пример дешифрования и отмечается важное качество языковых сообщений, а именно инвариантность языковых характери-

стик на разных уровнях, позволяющая найти хотя бы самые общие свойства кода. «Можно было догадаться, что при расшифровке использовалась частотность букв и комбинаций букв... которые почти не зависят от того, каким языком написан зашифрованный текст. Сразу же возникает предположение, что на том этапе, когда человек изобретал и развивал языки, возникли некоторые их инвариантные свойства, которые опять-таки неточно, но до некоторой статистически полезной степени присущи всем языкам» [Уивер 1957].

Можно считать, что на становление и развитие машинного перевода повлияли следующие этапы криптоанализа, касающиеся различных уровней языкового описания:

- исследование чисто внешних характеристик сообщения, таких, как частота отдельных элементов и сочетаний элементов;
- исследование дистрибуции элементов, составляющих текст;
- анализ билингв с целью выявления совпадающих элементов и использования некоторых ключевых элементов (собственных имен и т. п.) с учетом статистических и дистрибутивных характеристик;
- нахождение ключа и применение его ко всей последовательно сти закодированных сообщений [Kahn 1968; см. также: Кан 2000].

Основные вехи дальнейшего развития машинного перевода следующие: в 1947 г. А. Д. Бут и Д. Г. Бриттен разработали подробный «код» для пословного перевода. В 1948 г. Р. Г. Риченс предложил правила разбиения словоформ на основу и окончания. Первая конференция по машинному переводу прошла в Массачусетском технологическом институте в 1952 г. В 1954 г. в Нью-Йорке состоялся Джорджтаунский эксперимент, когда на машине IBM701 было переведено несколько специально подобранных русских предложений.

Основные этапы развития машинного перевода

В 1983 г. [Марчук 1983] было выделено три периода развития машинного перевода, каждый из которых охватывает примерно десять лет, а именно: 1946 — 1957 гг.; 1957 — 1967 гг.; 1967 — 1980-е г. Сейчас можно выделить и современный период, начавшийся примерно в середине восьмидесятых годов прошлого столетия и продолжающийся до настоящего времени. Основные общие характеристики этих периодов следующие.

Первый период включает появление идеи машинного перевода с помощью вычислительных машин, которые тогда были новым достижением, и связанный с этим энтузиазм. Оптимистические оценки ближайших перспектив машинного перевода базировались в значительной мере на таких же оптимистических воззрениях на перспективы развития электронной вычислительной техники. Казалось, что совсем близко создание искусственного интеллекта. Еще менее ясности было в вопросе о возможностях и границах формализации

применительно к такому сложному общественному интеллектуальному феномену, каким является естественный язык. Возникли многочисленные научные коллективы, многие из которых ставили своей целью быстрое получение машинного перевода с многих языков на многие другие посредством формализации смысла. Статистические и теоретико-информационные измерения языка привели к созданию концепции языка как кода. Внимание научной общественности к машинному переводу было привлечено первыми опытами, такими, как Джорджтаунский эксперимент. Особое внимание отводилось построению машинных словарей, хотя уже складывались представления об алгоритмах морфологического, синтаксического, лексического анализа для целей машинного перевода.

Основные итоги этого периода можно было бы суммировать следующим образом:

- выдвинута концепция языка как кода;
- созданы первые машинные словари экспериментального характера;
- разработано общее представление о машинном переводе, включающее машинный словарь, алгоритмы анализа и синтеза текста, программное обеспечение;
- созданы и опробованы первые экспериментальные системы машинного перевода, которые, хотя и носили рекламный характер, обладали рядом существенных для машинного перевода черт.

Наиболее плодотворным можно считать следующий, **второй период** в истории машинного перевода (1957—1967). Он закончился в некотором смысле декларированным отходом от машинного перевода как практической проблемы, но в целом характеризовался повышенным вниманием к нему как к теории. В этот период продолжались работы по созданию систем машинного перевода как в СССР, так и за рубежом. Однако основное внимание стала привлекать теория машинного перевода, поскольку стало ясно, что традиционная лингвистика не может дать четкие ответы на целый ряд проблем, связанных с формализацией языка и алгоритмизацией работы с ним. Слишком многие аспекты языковой деятельности оказались чрезвычайно далеки от формализации, что препятствовало работе с ними в режиме компьютерных программ. Можно утверждать, что именно в хронологических рамках второго периода началась реальная дивергенция научных интересов, впоследствии приведшая к тому, что машинный перевод стал представлять собой конгломерат лингвистических наук и не только лингвистического цикла, но и связанных с ЭВМ другими аспектами.

Здесь особо следует отметить концепцию языка-посредника. Хотя эта идея не оказала сколько-нибудь существенного влияния на создание промышленных систем машинного перевода, она стимулировала развитие сравнительно-исторического языкознания, исследований в области морфологии, синтаксиса, семантики, привела к соз-

данию теории семантических полей [Караулов 1976], заинтересовала математиков и логиков. Кульминацией этого периода можно считать доклад комиссии Национального научного фонда США, специально посвященный экономической целесообразности машинного перевода и перспективам научных исследований в этой области [Язык и машины 1968]. В этом докладе, исходя из реальной ситуации с переводами в США, делался вывод о невыгодности машинного перевода, недостижимости поставленных в нем целей в течение сколько-нибудь разумного времени, содержались рекомендации изучать иностранные языки вместо того, чтобы поручать дело перевода машине. Пессимизм доклада не получил поддержки многих ученых в США и за рубежом, однако он оказал серьезное влияние на оценку машинного перевода и повлек за собой уменьшение финансирования научных исследований.

Тем не менее за этот период был сделан существенный вклад в решение проблематики машинного перевода и, что не менее важно, смежных наук, взаимодействующих с переводом, как машинным, так и «человеческим». Так, можно отметить, что за этот период:

- осознана недостаточность исходных представлений о языке для его полного формального описания, необходимого для создания алгоритмов и систем машинного перевода;

- в плане развития концепции языка как кода были разработаны статистические и теоретико-информационные методы измерения речи;

- возникла идея языка-посредника;

- возникла и утвердилась идея о целесообразности отделения описания языка формальными методами от собственно алгоритма работы с этими описаниями;

- центр внимания исследователей переместился от лексики к синтаксису и другим уровням языковой системы;

- сформулированы основные идеи об использовании семантики для алгоритмов машинного перевода;

- созданы и опробованы в эксперименте разнообразные алгоритмы синтаксического и семантико-синтаксического анализа;

- выдвинута концепция синтеза (анализ через синтез);

- автоматизирован морфологический анализ для многих языков, т. е. были достаточно подробно разработаны способы анализа словоформ аналитических языков и созданы соответствующие алгоритмы и программы;

- начато составление обширных словарей для машинного перевода с разнообразной информацией;

- продолжена эксплуатация экспериментальных систем и накоплен важный опыт их программной реализации;

- уточнены требования к математическому и программному обеспечению систем и отдельных алгоритмов анализа и синтеза текстов;

– составлено большое количество частотных словарей, отработана методика анализа с использованием статистических и теоретико-информационных характеристик, сложились основы количественной лингвистики, статистической лексикографии и других наук статистического цикла;

– сделаны успешные попытки промышленной эксплуатации систем машинного перевода в режиме реального информационного обслуживания для некоторых видов текстов;

– отработана методика контекстного анализа для функционального разрешения основных лингвистических проблем.

Параллельно с развитием машинного перевода в этот период возникла идея создания в помощь человеку-переводчику автоматизированного словаря, получившего практическое применение.

В течение этого периода ускоренно развивалась вычислительная техника, появлялись новые возможности машинной реализации лингвистических формализмов, усилилось внимание к теоретическим проблемам лингвистики с точки зрения новых возможностей обработки текстов.

Третий период (1967 — 1980) характерен возникновением нового социального статуса машинного перевода. В это время образовалась наука о научно-технической информации, которая в то время называлась информатикой.

Слово «информатика» изначально в русском языке понималось как научно-техническая информация, точнее, информирование, и лишь в последние годы за этим термином закрепилось значение «компьютерная технология, умение работать с компьютером, умение программировать». В течение же третьего периода возникли информационные системы, базирующиеся на новых информационных технологиях, утвердились понятия автоматизированного информационного поиска, системы автоматического проектирования и другие интеллектуальные системы.

Если во втором периоде развитие научной мысли шло в некотором отрыве и с учетом новых возможностей вычислительной техники, то третий период, как можно считать, соединил новые достижения технической кибернетики с новыми результатами теории и практики лингвистических исследований в автоматической обработке текстов. Это слияние произошло на базе осознания социальных потребностей в машинном переводе — необходимости эффективного преодоления языковых барьеров. Произошел возврат к построению цельных систем машинного перевода. Однако этот процесс происходил достаточно медленно, хотя в это время и в СССР развернулись работы по созданию систем машинного перевода для практического применения. Можно при этом сослаться на разработку И. К. Бельской, которая в сотрудничестве с математиком Д. Ю. Пановым внесла выдающийся вклад в теорию и практику машинного перевода [*Бельская* 1969].

Рассмотрим положение с машинным переводом в настоящем, **четвертом периоде**, который, как можно считать, начинается примерно с 1980 г. В этом мы будем опираться на результаты детального исследования, проведенного П. Н. Хроменковым в его учебном пособии «Современные системы машинного перевода», опубликованном в 2005 г. [Хроменков 2005].

П. Н. Хроменков констатирует, что за последнее десятилетие значительно возрос интерес к использованию систем машинного перевода (СМП). Это положение вещей связано в первую очередь с бурным ростом и развитием вычислительной техники и постепенной интеграцией всемирного информационного пространства в единую сеть, прототипом которой является Интернет. Многие СМП, которые ранее являлись исключительно промышленными и работали на конкретных потребителей, получили в последнее время свое коммерческое развитие и вышли на рынок информационных технологий в качестве полноценного программного продукта, обладающего свойствами товара.

Сеть Интернет в этом столетии станет одним из основных средств общения и коммуникации среди как представителей бизнеса, так и обычных пользователей. При таком глобальном распространении «всемирной паутины», которое наблюдается в последнее время, число языков, на которых составляются документы и странички пользователей, будет постепенно приближаться к общему числу языков, известных в мире. По статистике, только для 6 % населения Земли английский язык является родным, при этом до настоящего времени 90 % всей информации в Интернете было представлено только на английском. Сейчас наметилась обратная тенденция, поскольку, как мы уже говорили, большинство населения мира предпочитает общаться на своих родных языках. Число неанглоязычных сайтов, по данным компании Global Reach, как пишет П. Н. Хроменков, также растет и уже в 2000 г. превысило число узлов на английском языке. 22,4 % информации в Интернете представлено на испанском, 14 % — на немецком, 12,3 % — на японском и 10 % — на французском языках. Число сайтов на русском языке также расширяется. К 2007 г. доминирующим языком сети Интернет должен был (по прогнозам) стать китайский.

В результате увеличивающегося интереса и возрастающих потребностей в осуществлении перевода больших массивов информации, на крупных узлах и в информационных службах приступили к внедрению технологии машинного перевода. Данные технологии позволяют осуществлять перевод в режиме реального времени. Такие системы применяются в основном в межнациональных многоязычных интра- и экстрасетях. Процесс создания сайтов, на которых можно получить перевод документа или странички, идет полным ходом. В Интернете насчитывается уже более 30 сайтов, где вам предлагают за невысокую плату или совсем бесплатно выполнить машинный перевод материалов. Уже более пяти лет на таком крупном и извест-

ном поисковом узле, как Alta Vista, используется машинный перевод в режиме реального времени.

Из хорошо известных систем можно упомянуть систему SYSTRAN, которая осуществляет перевод с английского языка на немецкий, французский, итальянский, португальский, русский, китайский, корейский, японский, испанский языки и обратно. Эта система может обслуживать несколько сотен запросов в день и пользуется большим успехом. Уже с 1994 г. в службе CompuServe проводится эксперимент по внедрению и использованию СМП в онлайн-форумах. Функция машинного перевода внедрена на 6 форумах. Для этих целей используется СМП Transcend фирмы Transparent Language, которая достаточно популярна среди участников данных форумов. Стоимость машинного перевода составляет при этом 1 цент за слово. По сравнению с компаниями, использующими труд людей-переводчиков, такая стоимость значительно ниже, поскольку перевод, сделанный человеком, обходится примерно в 25 центов за слово. Качество перевода человека может быть выше, однако для того, чтобы принять решение, какие документы требуют качественного перевода, вполне можно воспользоваться услугами СМП. В Интернете имеются и многие другие системы реального пользования.

Подробный обзор практически всех многочисленных систем, имеющих коммерческое применение, содержится в упомянутой работе П. Н. Хроменкова.

Глава 6

О ТЕОРИИ И ПРАКТИКЕ МАШИННОГО ПЕРЕВОДА

Двенадцать проблем машинного перевода

Машинный перевод затрагивает ряд проблем, которые продолжают существовать, несмотря на то что уже достаточно долгое время предпринимаются попытки их решения, более или менее удачные. Эти проблемы были в основной части сформулированы Кристианом Буатэ, руководителем Гренобльского центра машинного перевода во Франции, в его докладе на международной конференции по машинному переводу в Малайзии [Boitet 1991]. Рассмотрим эти проблемы в свете современных достижений (или при отсутствии таковых).

К. Буатэ выделяет три типа машинного перевода. Первый — это «информативный». Это грубый пословный перевод, ценность которого заключается в том, что он обеспечивает некоторое поверхностное знакомство с содержанием текста на незнакомом языке. Такие системы существуют для многих языковых пар, как показано в материалах П. Н. Хроменкова. Второй тип — «профессиональный». Качество перевода здесь сравнимо с качеством «человеческого» варианта и при не-

большом редактировании является полностью удовлетворительным. Такой перевод выгоден для больших объемов текстов (свыше 10 000 страниц в год) и для однородных текстов. Третий тип — «персональный перевод». Авторы подлежащих переводу текстов заранее избавляют их от неоднозначностей и работают в режиме диалога с компьютером. При этом появляется возможность улучшить качество перевода и достичь наибольшего удобства работы с ним. С такой классификацией систем машинного перевода можно согласиться и в настоящее время.

К. Буатэ далее утверждает, что машинный перевод есть не только область приложения идей вычислительной лингвистики, но и самостоятельная область знания. В ней можно выделить двенадцать основных проблем, образующих четыре группы: концептуальные проблемы, проблемы архитектуры, технические и инженерные.

Рассматривая **концептуальные проблемы**, автор выделяет среди них следующие: каким образом представлять и разрешать неоднозначности разного рода? Каждой единице перевода, как правило, сопоставляется несколько возможных абстрактных представлений. Надо не только определить нужное в данном тексте, но и предвидеть появление других неоднозначностей вследствие данного выбора. Для решения этого вопроса целесообразно ввести соответствующее «неоднозначное программирование», предусматривающее разные возможности разрешения неоднозначности.

Комментируя такую формулировку концептуальных проблем, можно отметить, что выбор одного из вариантов «абстрактного» представления неоднозначности действительно представляет трудность, тем более если поставить задачу определения других возможных неоднозначностей. Однако «неоднозначное программирование» — слишком общий и не совсем ясный термин. Его можно понимать как необходимость создания нескольких вариантов алгоритма перевода неоднозначной входной текстовой единицы.

Вторая проблема — переводные соответствия. Как обеспечить точное соответствие между абстрактной структурой и текстом? Конкретное высказывание всегда отличается от его абстрактной модели. Такая проблема действительно существует, но только при том, что для перевода применяются некоторые абстрактные модели. Следует напомнить, что гренобльский подход, в становлении которого на ранних стадиях разработки принимал участие математик Б. Вокуа, базировался на возможности применения универсального языка посредника — *langue pivot*, — которого придерживались в машинном переводе многие математики.

Концептуальной проблемой является также использование формализмов. Каким образом создать эффективные средства анализа из чистых формализмов, добавляя к ним процедурные или эвристические знания? Это приходится делать, утверждает К. Буатэ, поскольку ни одна из действующих систем машинного перевода не использует только декларативные знания.

Здесь можно сделать некоторое пояснение. Многие математики высказывали точку зрения, что необходимо разделение труда между математиками и лингвистами в создании СМП. Дело лингвистов — дать описание языковых фактов, устройства языка, т. е. предоставить математикам лингвистическое статическое описание языка, того, как он устроен. Создание алгоритма работы с лингвистическим описанием для перевода — дело математиков, поскольку алгоритм — понятие математическое. Однако такой подход себя не оправдал, так как для собственно перевода нужны знания, которые называются процедурными. Эти знания основаны на описании некоторых процедур, необходимых для определения тех или иных лингвистических представлений или единиц. Например, в морфологии значение тех или иных суффиксов или окончаний зависит от того, как меняется данное слово, в каком оно стоит падеже или иной грамматической категории. Поэтому базироваться только на декларативных знаниях при построении алгоритмов оказалось невозможным, что и констатирует К. Буатэ.

Среди **проблем архитектуры** наиболее важные относятся к неполноте описания языка в целом. Как построить и использовать грамматические формализмы, которые дали бы возможность описать разные варианты языка? Как ведут себя «микроязыки» или «подъязыки» в рамках единой СМП? Как настроить систему на конкретный подъязык? Как конструировать и использовать грамматические формализмы, которые позволили бы описать варианты (микроязыки или подъязыки) в рамках единого языка, приспособив конкретную систему машинного перевода для данного языка во всей его общности к конкретному подъязыку посредством настройки? Главным в этом вопросе является, по мнению К. Буатэ, лексический аспект.

Следующая проблема того, что К. Буатэ относит к «архитектуре», — обучение на основе текстов. Возможно ли создать системы, которые реально учились бы на основе существующих параллельных текстов и на корректировке ошибок, отмеченных профессиональными редакторами? Статистически основанные СМП могут быть эффективными для перевода коротких сообщений. Создание массивов параллельных текстов также может быть выгодным для разработки таких систем. Что касается грамматических правил, то их гораздо труднее выводить автоматически из сравнения параллельных текстов.

«Степень понимания». Как воспроизвести «нечеткое» понимание человека-переводчика? Как воспроизвести высказывания частично в *content-bound way*, а частично в *structure-bound way*? Основанный на знаниях машинный перевод возможен в некоторых контекстах, где можно положиться на полное знание и полное представление некоторой предметной области или задачи. Например, можно переводить с помощью компьютера инструкции по уходу за ядерными реакторами. Однако во многих случаях даже в рамках конкретной предметной области возможен отход от чисто предметного содержания. Приводится такой пример: *I would prefer the meeting to take place*

downtown, because I will be with my wife... She wants to visit her niece, who is studying law there. В качестве рецепта для преодоления такого рода трудностей, когда имеет место отход от предметной области, К. Буатэ предлагает научить машину истолковывать высказывания частично по содержанию, частично по структуре, а затем гармонично их соединять. Следует отметить, что не совсем понятно, как это можно реально сделать, так что проблема остается. Она носит как лексический, так и грамматический характер.

«Технические проблемы», оценка качества и перспективы

Технические проблемы заключаются в организации многоязычных лексических баз данных. Как организовать большие многоязычные, многоцелевые, многотеоретические базы лексических данных, на основе которых можно было бы строить автоматически конкретные словари для машинного перевода? Как работать с частичным или зашумленным вводом? Все современные системы построены с учетом того, что введенный для перевода текст не содержит ошибок. Однако, как правило, в нем встречаются не только искажения знаков, но и ошибки разного рода: грамматические, эллиптические конструкции, нарушающие синтаксические правила, небуквенные графемы и пр. Коррекцию на уровне знаков текста можно осуществить, опираясь на статистические закономерности (см., например, кандидатскую диссертацию Л. Ю. Коростелева [Коростелев 1985]). Следующая проблема — как работать с целым текстом, а не с отдельными фразами. Эта задача не является технической, поскольку ее решение требует, по крайней мере, алгоритмического установления анафорических связей, обращения к фрагментам предыдущего и последующего текстов для разрешения лексической многозначности и т. п.

Следующий класс проблем К. Буатэ называет **инженерными**, хотя они также являются достаточно сложными и существенно затрагивают лингвистику. Первая из них: как кодировать многоязычные тексты? Так, например, латинский алфавит, используемый в английском, французском, испанском и других языках, имеет диакритики и двухбуквенные сочетания, которые означают разные звуки и поэтому вызывают трудности в автоматическом распознавании. Вторая проблема: как увязывать «архитектуру» с применениями? Для разных применений предлагаются разные решения: модели трансфера, перевода с использованием смысла, пословный и буквальный перевод. Поэтому важная проблема — дать четкую картину того, что можно ожидать и чего нельзя ожидать от перевода. И наконец, последняя проблема, также называемая инженерной. Как измерить прогресс в машинном переводе и вклад исследований по такому переводу в фундаментальные представления о характере языка и его использовании?

К. Буатэ заключает, что, наподобие ядерной реакции, машинный перевод является мечтой нашего века. В противоположность термо-ядерной реакции он уже дает результаты, однако финансируется далеко не так. В чем дело? Возможно, отсутствует истинно научный элемент, а именно измерение прогресса. Машинный перевод и связанная с ним лингвистика, видимо, не вносят еще какой-либо измеримый, фундаментальный вклад в науку.

Даже меньшую цель, а именно сравнение существующих СМП, трудно достичь. В контексте информативного перевода может быть организована конкуренция между системами, работающими в одинаковых языковых парах и в тех же тематических областях. Однако это пока что сделано только в отношении японо-английского перевода. Что касается перевода профессионального, это почти невозможно: система может показать свою ценность только после настройки на конкретную область (по терминологии) и на стиль (по грамматике). Даже если подобное сравнение организовано на справедливой основе, т.е. финансированием нескольких исследовательских групп, пока нет согласия относительно критериев, которые нужно при этом применять. Суждения о качестве весьма субъективны. Измерения стоимости и эффективности более объективны, но требуют много времени для того, чтобы пользователь привык к системе. Важным критерием при этом является легкость улучшения системы — пополнения словарей, улучшения эффективности работы алгоритмов и пр.

Таким образом, несмотря на то что некоторые из этих проблем получили приемлемые для практики решения, улучшение качества собственно машинного продукта при переводе требует учета перечисленных теоретических и практических моментов, сформулированных уже достаточно давно. Причина такого положения заключается прежде всего в том, что перевод, как было уже констатировано, является сложной проблемой с самыми разнообразными аспектами, и поэтому его моделирование для осуществления на компьютере требует, по существу, разработки искусственного интеллекта в довольно объемной области умственного усилия.

Глава 7

ЭФФЕКТИВНОСТЬ МОДЕЛИРОВАНИЯ ПЕРЕВОДА И ПОНЯТИЕ «ПОДЪЯЗЫК»

Естественный язык и «подъязыки»

В языковой практике хорошо известно, что любой естественный язык как некоторая реально существующая система может рассматриваться как состоящая из некоторых более мелких подсистем. По-

следнее время большое распространение получили так называемые языки профессионального общения: делового, коммерческого, учебного и т. п. (см., например, материалы III Международной научно-практической конференции «Актуальные проблемы лингвистики и лингводидактики иностранного языка делового и профессионального общения», прошедшей в 2008 г. в Российском Университете Дружбы народов [Актуальные проблемы лингвистики и лингводидактики иностранного языка делового и профессионального общения 2008]).

В докладе Е. Н. Малюги на этой конференции совершенно справедливо отмечается, что в настоящее время всякая профессиональная деятельность так или иначе связана со знанием иностранных языков. Специалист (мы говорим о специалистах с высшим образованием или, по крайней мере, со средним специальным), чтобы преуспеть в своей профессии, должен быть знакомым с достижениями в соответствующей предметной области, имеющими место в других странах и описанными на иностранных языках. Таким образом, повышение его компетентности связано с углублением его знаний определенных иностранных языков. Профессионально-ориентированный подход к обучению иностранным языкам в неязыковых вузах связан с формированием у студентов способности иноязычного общения в специальных, деловых и научных сферах на основе профессионального мышления. В настоящее время, по сравнению с прошлыми годами, постоянно растет потребность в развитии языковой составляющей вузовского образования, формируемой в контексте глобализации на основе устной и письменной коммуникации, а также лингвокультурных аспектов. Потребность в иностранных языках возникает в процессе использования телекоммуникационных технологий, для обработки больших потоков информации, в связи с развивающимися международными отношениями в сфере профессиональной деятельности специалистов [Марчук 2008].

Понятие «подъязык» достаточно широко. Мы не будем специально останавливаться на рассмотрении разных аспектов этого понятия, отметим только, что часто оно взаимодействует с понятием «стиль».

Системность в лексике

Есть ли вообще системность в лексике и насколько правомерно и эффективно понятие «подъязык»?

В книге О. А. Корнилова [Корнилов 2003] этот вопрос рассмотрен достаточно подробно. Приведем основные точки зрения, базируясь на его исследовании.

Вопрос о системности лексики можно считать риторическим. Лексическая системность может считаться противоречивой, непоследовательной и т. п. «Язык — весьма сложная система, в которой не толь-

ко все «системно», но в которой нередко наблюдаются противоположные тенденции, постоянно осложняющие эту систему» ([Будагов 1976; цит. по: Корнилов 2003]. Зд. и далее цитаты из других авторов приводятся по данным О. А. Корнилова). «Наши языки — это системы, которые по существу своему удивительно асистемны... Человеческий язык одновременно и система, и антисистема» [Будагов 1976]. В обобщенном виде, пишет О. А. Корнилов, аргументы против признания системного характера лексики можно сформулировать следующим образом:

- системность лексики — простое отношение системных связей и отношений действительности; в силу своей внелингвистической природы эта системность не может быть предметом лингвистического описания;

- словарный состав подвижен, практически безграничен и в силу этого не поддается систематизации;

- многозначность, порождаемая семантическим варьированием слов, создает множественность связей; эти связи накладываются друг на друга, пересекаются, что создает трудности для элементарной группировки лексических единиц;

- в лексике очень много периферийных, окказиональных явлений и исключений.

Среди положительных взглядов на системность в лексике отмечаются три направления:

- совокупность понятийных полей, которым в языке соответствуют языковые поля, где каждое слово покрывает соответствующее понятие;

- совокупность лексико-грамматических групп слов;

- совокупность лексико-синтаксических групп слов.

В отечественном языкознании центральное место в лексико-семантической системе занимает слово. Принципиально важно, что для отечественного языкознания характерно целостное понимание лексической системы, учитывающее и развивающее два аспекта системы: **внутренний** (по отношению к центральной единице системы — слову) и **внешний**. В своей работе [Марчук 1992] автор обосновал точку зрения на системность в лексике, обозначив два принципиально разных подхода к описанию лексики — текстоцентрический, который может примерно соответствовать внешнему аспекту описания лексики, и словарноцентрический, который может соответствовать внутренней системности лексики.

Интересно, что в таких науках, как теория коммуникации, вопрос о системности лексики либо не ставится вообще, либо в качестве элементов лексической системы представляются виды коммуникации. Причем они никак не определяются в качестве именно элементов лексического состава языка, а привязываются и рассматриваются с точки зрения разных видов коммуникации, как вербальной, так и невербальной (см., например, [Садохин 2004, Почепцов 2003]).

Так, в последней работе раскрываются общие характеристики коммуникации, социологические, психологические и семиотические модели коммуникации, нейролингвистическое программирование и т. п., но лексический аспект этих моделей и типов коммуникации не выделяется каким-либо специальным образом.

Однако в целях автоматической обработки текстов на естественных языках понятие «подъязыка» или каких-либо других сходных по функциям конструкций, так или иначе систематизирующих лексику, является необходимым. Поэтому рассмотрим один из практических способов решения такой задачи.

Следует добавить также, что понятие системности в языке, наличие или отсутствие в нем делений, более мелких, чем язык в целом, необходимы также и для перевода как для комплексной лингвистической задачи, тем более для перевода, который осуществляется с помощью современных информационных технологий.

Функционально-коммуникативная модель микроподъязыка

Потребность в точном, однозначном, формальном описании языка и его компонентов, в таком описании, которое можно было бы получить, исследуя речь как результат работы языка, особенно проявляется в вопросах автоматической обработки текста. Компьютер не может абстрактно мыслить, делать формальные заключения на основе обобщения неформальных описаний, самостоятельно преобразовывать декларативные знания в процедурные, как бы этого ни хотелось сторонникам математико-логического направления в прикладном языкознании.

Рассмотрим одну из моделей, примененных в машинном переводе, с точки зрения ее принципиальной структуры и результатов. Эта модель осуществлена в практической разработке и промышленном применении одной из систем машинного перевода. Представляется интересным также, что объектом исследования и разработки стал такой своеобразный язык, как китайский. Подробно подход описан в кандидатской диссертации В. М. Зелко [Зелко 1991].

Китайский язык с общелингвистической точки зрения, как и любой другой естественный язык, можно рассматривать не как монолитную систему, а как некоторую суперсистему языковых подсистем, функционирующих в тесном единстве друг с другом. В результате такого функционирования порождается множество более мелких, разнообразных конкретных языковых микросистем, которые служат для осуществления актов коммуникации в конкретных областях человеческой деятельности. В. М. Зелко следующим образом представил структуру китайского языка как набора более мелких составляющих его элементов (рис. 7).

КИТАЙСКИЙ ЯЗЫК
(суперсистема языковых подсистем)

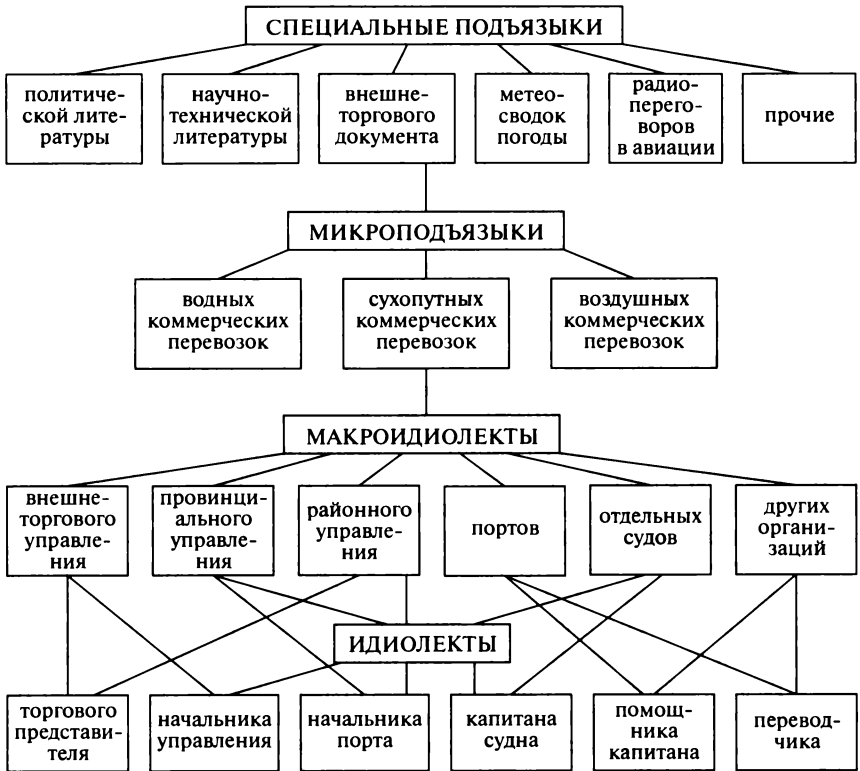


Рис. 7. Структура естественного языка (на примере китайского подъязыка)

Ясно, что возможны и другие варианты классификации составляющих естественного языка. Повторяем, что здесь мы имеем практическую разработку для машинного перевода достаточно четко определенного микроподъязыка водных коммерческих перевозок.

Специальные языки можно определить как класс языковых подсистем, обслуживающих определенные сферы человеческой деятельности, создаваемые и функционирующие в пределах определенных коллективов и групп, обеспечивающие фиксацию, хранение, передачу и обработку нового знания или необходимой специальной информации. Специальные языки ограничены в лексическом составе, отличаются своей терминологией, особенностями стиля и грамматического строя. Они образуют иерархическую структуру, выступая как макроязыки по отношению к входящим в них малым подъязыкам, которые характеризуются наибольшей полнотой, конечностью в смысле ограниченности лексического состава и других языковых

характеристик и являются наиболее формально ограниченными. Эти более мелкие лингвистические образования можно назвать **микроподъязыками**.

В данной работе В. М. Зелко, исходя из описанной структуры поступающих на анализ текстов, микроподъязыками внешнеторгового документооборота считаются такие лингвистические микросистемы, которые, включая ограниченное множество синтаксических состояний, обладают относительной коммуникативностью и в состоянии своими собственными средствами описать любой процесс (действие, факт, состояние) в узконаправленной предметной области.

В свою очередь, микроподъязыки выступают главными в некоторой следующей иерархии. В рамках каждого микроподъязыка можно выделить некоторый набор составляющих его *макроидиолектов*.

Макроидиолект можно рассматривать как набор языковых элементов и их отношений в текстах с однородной тематикой. Применительно к документам водных коммерческих перевозок макроидиолект есть совокупность лингвистических единиц определенной специальной направленности, используемых отдельной организацией (управлением внешнеторговых перевозок, районным управлением и т. п.) для обеспечения акта коммуникации при выполнении производственных и других задач.

В свою очередь, макроидиолекты состоят из более мелких языковых элементов, которые можно назвать *идиолектами*.

Идиолект — совокупность лингвистических компонентов, используемых конкретным лицом для поддержания процессов общения в рамках данного конкретного подъязыка. Таким лицом можно считать торгового представителя, капитана судна и др.

Каждый человек владеет множеством различных диалектов, относящихся не только к конкретным подъязыкам, но также и к разным подсистемам и целым языковым системам (например, иностранными языками). Это характерно для любого человека, живущего в обществе. Однако в рамках узкой направленности своей производственной деятельности каждый индивидуум пользуется ограниченным набором языковых средств, которые представляют собой совокупность отдельных идиолектов. Чем уже аспект деятельности человека, т. е. чем более узка предметная область, описываемая языком, тем меньше требуется языковых компонентов для поддержания общения в строго ограниченных рамках данного процесса.

Лингвистическое наполнение конкретных идиолектов, используемых отдельным лицом, а также их количество и разнообразие зависят от уровня образования, характера деятельности, занимаемой должности, социальных условий и многих других аспектов. Этим, в частности, можно объяснить некоторые различия в терминологическом, грамматическом и стилистическом подходах к составлению идентичных по содержанию документов, авторами которых являются разные лица.

В. М. Зелко далее выражает мнение, что теоретически возможно, зная характерные особенности идиолекта конкретного лица, определить составителя документа, хотя практически эта задача остается нерешенной. Однако опыт перевода документов показывает, что в целом ряде случаев какие-то стилистические, грамматические, лексические особенности использования языка позволяют практически безошибочно определить автора текста. Другое дело, что иногда эти в значительной мере интуитивно устанавливаемые признаки достаточно трудно сформулировать в виде конечного числа строго детерминированных правил. Тем не менее в конкретных видах дискурса и эта задача может быть решена.

Возвращаясь к рассмотрению проведенного исследования, можно отметить, что на уровне макроидиолектов задачу атрибуции текстов можно решить. Макроидиолект как система мелких идиолектов дает возможность, по выражению В. М. Зелко, «читать между строк» и выявлять такую информацию, о которой прямо в документе не сказано, но которая значительно дополняет документ, что весьма важно для достижения адекватности перевода. Например, терминологический анализ отдельного макроидиолекта за определенный период времени может указывать на изменение организационной структуры и подчинения, корректировку производственных и других задач и т. п. С помощью стилистического и грамматического анализа можно выявить замену лиц руководящего состава, новые пополнения на судах, в районных управлениях, повышение уровня профессиональной подготовки лиц, представляющих данную организацию, и т. п.

Рассмотрим этапы построения функционально-коммуникативной модели.

Объектом исследования явились оригинальные материалы, разновидность внешнеторгового документооборота по линии советско-китайской торговли. Были проведены два вида анализа: структурно-номенклатурный и лексико-грамматический.

Структурно-номенклатурный анализ проводился ручным способом, с частичным привлечением вычислительной техники, методом отбора, сортировки и инвентаризации отдельных единиц входного текста с их последующей систематизацией и группировкой по функционально-коммуникативным признакам. В результате выявлены и изучены особенности структуры документов, определены их типы и виды в зависимости от тех или иных элементов формализации.

Было установлено, что структура документа включает в себя четыре основных компонента: заголовок, адрес, текст и подпись.

Заголовок состоит из номера документа, даты и времени его составления, серийности или вида документа, регистрационного номера и других атрибутов служебного делопроизводства.

Адрес, или адресная часть документа, включает в себя наименования различных органов управления, названия судов, портов,

диспетчерских, должности и фамилии тех лиц, кому адресован конкретный документ.

Текст — наиболее полная и важная часть документа. Для него характерны многие знаки рубрикации и условной формализации (пункты, подпункты, круглые скобки, косая черта, цифры в натуральном виде, буквы латинского и русского алфавитов и пр.). Следует отметить, что такого рода знаки оказывают существенное воздействие на формальный анализ текста. Это влияние носит двойственный характер: так, скобки с учетом содержащегося в них отрезка текста могут означать границы простого предложения; тире, апостроф, точка после буквы инициала могут означать некоторые промежуточные границы анализа. Вместе с тем эти же знаки могут и затруднять выбор правильного решения относительно синтаксической структуры или морфологического состава анализируемой языковой единицы.

Подпись указывает на должностное лицо или орган, составивший и подписавший данный документ. Иногда в документе проставляется также дата и время его составления и подписи.

С точки зрения номенклатурной организации по результатам анализа весь служебный документооборот можно поделить на три основные группы в зависимости от того, кто является составителем документа и кому он адресован, т. е. по направлению документообмена, и на пять основных групп в зависимости от характера сообщаемой информации.

По направлению документооборота

1. **Сверху вниз**: от вышестоящей организации к нижестоящей. Например, из провинциального управления в районное.

2. **Снизу вверх**: от нижестоящей организации к вышестоящей. Например, от судна в районное управление, из районного в провинциальное и т. п.

3. **По горизонтали**: между равноправными инстанциями. Например, от одного судна к другому, из одного районного управления в другие и т. п.

По характеру сообщаемой информации

1. **Указания**. Основная коммуникативная функция этого вида документов состоит в том, чтобы довести до исполнителей решения органов управления или соответствующих начальников, а также сообщить о поставленных задачах. К данному виду документов относятся различного рода приказы, распоряжения, указания руководства и пр. Этот класс документов представляет наибольший информационный интерес, так как отображает в полной мере состояние и перспективы развития двустороннего товарообмена и вообще какой-либо проблемы.

2. **Уведомления** предназначены для информирования подчиненных инстанций, руководства, а также взаимодействующих ор-

ганов о событиях, не носящих плановой регламентации хозяйственной или производственной деятельности. Они включают сообщения о мероприятиях политического характера, о чрезвычайных происшествиях, нарушениях существующего порядка и правил, проведении проверок, прибытии делегаций и пр. Документы-уведомления от высших органов к низшим тесным образом связаны с документами-указаниями, так как последние в большинстве случаев содержат элементы уведомления о каких-либо событиях, на основании чего даются соответствующие указания. Надо отметить, что при этом возможно появление в них самых различных лексических элементов, новой терминологии и т. п.

3. П л а н ы, т. е. всякого рода документы, имеющие отношение к вопросам планирования. Они содержат оценку обстановки, определяют замысел действия, планируют работу органов управления, его инстанций и исполнителей.

4. О т ч е т ы служат для информирования соответствующих инстанций о выполненной работе, о своей деятельности и состоянии, местонахождении, о характере деятельности сопредельной стороны. Эта категория документов в количественном отношении самая многочисленная благодаря сообщениям судов о своем местонахождении, которые, согласно подсчетам, составляют не менее 42 % от всего массива анализируемого материала.

5. З а п р о с ы даются органом управления или отдельным лицом для уяснения вопросов, связанных со всевозможной деятельностью организаций или отдельных лиц. Данные документы не носят планового характера. Они часто связаны с нарушениями установленного порядка документообмена, несвоевременным представлением плановых отчетов, нечетким или неполным пониманием предшествующего сообщения, а также с внезапно возникающей обстановкой, могущей привести к срыву планового задания.

Такого рода классификация является в какой-то степени условной, поскольку документы одного вида могут содержать элементы документов другого вида и разные виды могут пересекаться. Но тем не менее основная типологическая классификация помогает правильно определять основную сущность документа и выбирать способ его анализа и перевода.

Математический подсчет объема документооборота по характеру сообщаемой информации в рамках принятой классификации позволяет судить о степени представленности тех или иных документов в общем массиве анализируемых материалов. Так, установлено, что наибольший процент занимают отчеты — 62 % от общего объема, затем следуют уведомления — 16 % и планы — 14 %. Наименее часты указания — 3 % и запросы — 5 %. Данные о численном содержании разного вида документов дают возможность строить предположения о типичной картине информационного документооборота в определенных условиях коммуникации.

Из вышеприведенного можно сделать общий вывод о том, что структура информационного документообмена заранее предопределена и строго регламентирована социальной и технологической сферой общества, воздействие которой на языковые средства деловой документации определяет количественный и качественный состав лингвистического наполнения конкретного микроподъязыка.

Идея специального языка для машинного перевода практически используется в системе СИСТРАН, работающей в фирме «Ксерокс». Достаточно трудно создавать ограниченный язык для носителей данного естественного языка и доказывать необходимость пользоваться таким урезанным языком. Иностранцы воспринимают такую необходимость легче. Тем не менее опыт корпорации «Ксерокс», особенно в части обширных переводов материалов на экспорт, показал, что такое решение практически осуществимо, т. е. ограниченность подъязыка дает возможность организовать практически удовлетворительное применение машинного перевода или машины в помощь переводчику.

Наконец, ограничение семантики подъязыка позволяет строить эффективные механизмы типа «искусственный интеллект». Современное состояние машинного перевода таково, что прямое применение результатов систем искусственного интеллекта для организации на них систем анализа или синтеза невозможно, ибо «семантика» искусственного интеллекта несоизмеримо беднее любой, даже самой ограниченной семантики любого подъязыка естественного языка. Единственно мыслимым использованием результатов систем искусственного интеллекта может быть некоторая классификация текстов в смысле отнесения их к тому или иному подъязыку, с тем чтобы далее можно было бы выбрать тот или иной словарь.

Целью **лексико-грамматического анализа** является выделение как общезыковых закономерностей исследуемого массива текстов, так и их специфических лингвистических особенностей для создания воспроизводящей модели микроподъязыка водных коммерческих перевозок. Изучение лингвистических закономерностей проводилось в рамках проведенной структурно-номенклатурной классификации [Зелко 1991].

За отправную точку анализа было взято слово как важнейшая единица китайского языка. Известно, что есть много определений слова. На наш взгляд, заслуживает внимания его определение, существующее в области машинного анализа текстов, как цепочки букв, ограниченной двумя соседними пробелами или знаками препинания. Знак препинания тоже может быть словом, если взять, например, телеграфный дискурс. Так, в английском языке знаки препинания обозначаются следующими словами: *prd* (period), *pgh* (paragraph — обозначение абзаца), *cm* (comma — запятая), *dsh* (dash — тире) и пр. Кроме того, буква может быть обозначена специальным словом. В английском языке для того, чтобы буква, например, обозна-

чающая инициал собственного имени или часть какой-либо аббревиатуры, не была искажена или утрачена, существуют специальные обозначения букв алфавита: *a* — *able*, *b* — *baker*, *c* — *charlie*, *d* — *dog* и т. д.

Применительно к китайскому языку определение слова как совокупности букв (знаков) между соседними пробелами не очень удобно вследствие отсутствия видимых, четко обозначаемых на письме границ слов. Пробел между морфемами одного слова и между словами там одинаков. За видимую границу слова можно принять знаки пунктуации, однако этого недостаточно для словарной сегментации текстов. Для определения границ слова китайского языка должны быть использованы семантические, грамматические и статистические критерии. Тем не менее определение слова как совокупности знаков может быть использовано и для китайского языка. На этапе лингвостатистического исследования оно может быть положено в основу алгоритма компьютерного анализа заданного пакета текстов.

Мы подробно останавливаемся на описании решений в отношении китайского языка по той причине, чтобы проиллюстрировать, как особенности языковой структуры естественного языка влияют на состав и строение алгоритмов обработки текстов. Европейские и восточные языки в этом отношении обладают различной спецификой, и модели языков, отличающихся от европейских, должны учитывать совершенно иные языковые факторы.

Продолжим рассмотрение решений относительно китайского языка. В этом языке существуют односложные и многосложные слова. Под односложными понимаются слова, состоящие из одного слога, одной морфемы; это отдельная лексическая единица, передаваемая на письме одним иероглифом. Многосложное слово включает в себя две и более морфемы — иероглифа. Можно рассматривать это как закономерность всего китайского языка. При анализе исследуемого микроподязыка также выявлено незначительное количество односложных слов, обладающих определенным значением и синтаксической самостоятельностью. Например: (*ren*) — *человек*, (*ba*) — *восемь* и пр.

В количественном отношении односложные слова в исследуемом массиве представлены наиболее бедно в отличие от двусложных, занимающих преобладающее положение. Эта особенность данного микроподязыка соответствует общей закономерности китайского языка, где двусложное слово является статистически доминирующим речевым комплексом.

В целом на лексическом уровне исследуемый микроподязык характеризуется, с одной стороны, большой насыщенностью специальной терминологией, с другой — реализуя принцип экономии языковых средств — стремлением к использованию различного рода аббревиаций.

Терминология, относящаяся непосредственно к хозяйственной и производственной деятельности рассматриваемого ведомства, в се-

мантическом плане отвечает основным требованиям, предъявляемым к термину, — однозначности и краткости. Однозначность специальной терминологии тесно связана с многосложностью терминов, сложная структура которых сводит к минимуму любые проявления многозначности. При анализе текстов данного класса установлено, что чем больше составных элементов включает в себя термин, тем конкретнее становится его значение. Фактически требования семантической точности, предъявляемые к термину в рамках данного микроподъязыка, исключают двойное толкование специальной терминологии. Эта особенность способствует адекватной трансформации китайских терминов в русские эквиваленты, тем самым дает основание предполагать возможность разработки автоматизированных методов машинной атрибуции текстов исследуемого массива.

В отличие от специальной терминологии, общие лексические единицы исследуемого микроподъязыка не связаны жесткими ограничениями по однозначности и иногда могут принимать различные значения в зависимости от описываемой ситуации. Это касается главным образом глаголов, а также той лексики, которая характеризуется возможностью омонимии. Глагол, как часть речи, наряду с обозначением самого действия или состояния предмета в китайском языке может обозначать также цель или образ действия, а иногда и признак предмета. В этом заключается его основное отличие от глаголов европейских языков, что порождает ряд проблем при выборе эквивалента перевода. Китайские глаголы занимают большой пласт многозначной и безэквивалентной лексики, адекватный перевод которой представляет определенную трудность, а порой становится невозможным без экстралингвистической информации. Ряд глаголов обладает широкой сочетаемостью и допускает после себя дополнения, от значения которых зависит их переводной эквивалент.

В этом отношении можно также отметить особенности глагола и в других восточных языках, что необходимо с точки зрения моделирования перевода и составления соответствующих алгоритмов автоматического анализа текста и синтеза переводящих словоформ. Так, в кандидатской диссертации Али Реза Валипура [*Валипур* 1998] специально описываются глагольные формы и конструкции русского языка в сопоставлении с персидскими. Русский и персидский языки принадлежат к разным структурным типам, поэтому выявление общих черт и различий их морфологических систем представляет интерес в структурно-типологическом плане. По морфологическому строю русский язык относится к флективным языкам, а персидский — к аналитическим. Исследование А. Р. Валипура показало, что и в аналитических языках может наличествовать много флективных признаков, и, наоборот, во флективных — больше аналитических. Глагол — центральная категория морфологического строя языка, составляющая богатую и сложную систему форм, предикативный и модальный центры предложения. Составление эффективных алгоритмов

для автоматизированной обработки текста при машинном переводе требует осуществления сопоставительного анализа глагольных форм и конструкций в русском и персидском языках.

В диссертации А. Р. Валипура факты двух языков сгруппированы по трем признакам: сходные, тождественные в этих языках; характерные для обоих языков, но не тождественные; явления одного языка, отсутствующие в другом языке.

В персидском глаголе различаются морфологические категории лица и числа, времени, наклонения, залога. Эти категории находят свое выражение как в простых, так и в сложных формах глагола. И в русском, и в персидском языках глагол имеет одинаковое лексико-грамматическое значение: он обозначает действие, состояние, отношение как процесс (сравним с вышеприведенной квалификацией китайского глагола). Характерной особенностью русского глагола является наличие в нем категорий вида и рода. В персидском языке ведущей является грамматическая категория времени.

Таким образом, лексико-грамматический анализ в рамках определенного микроподъязыка должен базироваться на принципиальных отличиях в грамматических системах сравниваемых языков и учитывать их при составлении алгоритмов.

Вернемся к рассматриваемому микроподъязыку китайских документов водных коммерческих перевозок [Зелко 1991].

Перевод глаголов с широким семантическим значением должен осуществляться с учетом дополнений, образующих целый ряд устойчивых словосочетаний или языковых штампов. Учитывая относительную лексическую ограниченность данного подъязыка, выявление таких словосочетаний не представляет собой особо трудную задачу. Она может быть решена методом компьютерного анализа сочетаемости словоформ в текстах.

Проблема *омонимии*, т. е. принадлежности словоформы к разным частям речи, также представляет собой трудность для автоматической обработки текста. В разделах, посвященных анализу для машинного перевода с английского языка на русский, мы рассмотрим конкретные способы разрешения омонимии для англо-русской языковой пары. Эти способы безусловно могут быть применимы с соответствующими модификациями и для других языковых пар, и в других направлениях перевода. В рамках же рассматриваемого подъязыка китайских текстов проблема омонимии разрешается с учетом того, что в данном дискурсе существует всего несколько видов типовой омонимии. Частные случаи омонимии представляют в основном служебные слова, которые, несмотря на высокую частотность, сравнительно немногочисленны. Для решения этой проблемы нужно исходить из того, что любые слова-омонимы употребляются не изолированно, а в составе определенных групп (фрагментов, сегментов) текста.

Задача заключается в том, чтобы выявить эти группы и перевести слова-омонимы с учетом их контекстуального значения. Вследствие

ограниченности микроподъязыка количество употребления конкретного слова-омонима минимально. В среднем омоним употребляется в 3—4 вариантах в исследуемом массиве, редко отмечается до 10 вариантов. При разработке машинного словаря учитываются все контексты, в которых встречается данное слово-омоним, и в каждом конкретном контексте даются правильные решения относительно его реальной значимости.

Особенно остро стоит вопрос об анализе и переводе *аббревиатур*. Анализ текстов заданного класса (речь идет о рассматриваемом китайском микроподъязыке, но этот анализ показывает также и общие особенности аббревиации) позволяет выявить употребление аббревиатур с учетом сферы их применения. В. М. Зелко выделяет три типа аббревиации, исходя из особенностей исходного массива документов: общепринятая, контекстно-зависимая, ситуативно-обусловленная.

Общепринятая аббревиация не вызывает каких-либо трудностей при переводе. Относящиеся к ней термины являются общепризнанными, часто даются во всех общеупотребительных словарях, их можно встретить в широкой печати, в любой литературе и т. п. Проиллюстрировать этот тип аббревиации можно, например, такими понятиями, как РФ (Российская Федерация), ООН (Организация Объединенных Наций) и т. п.

Контекстно-зависимая аббревиация — это такой вид сокращений лексических единиц, который возможен только в условиях определенного контекста. При этом однозначное их понимание возможно только с учетом содержания текста или наличия в нем других лингвистических компонентов, раскрывающих значение аббревиатур. В качестве примера можно привести аббревиатуры, которые можно встретить в разделах рекламной печати типа «Экстра-М» или «Центр-плюс», например реклама: «Окна ПВХ», «Меняю кв. 2-х комн. на 3-х комн.» и т. п.

Третий вид аббревиации — ситуативно-обусловленная. Это высшая степень сокращения лексических единиц в результате общей компрессии текста документа под воздействием ранее созданной информационной ситуации. Факт применения ситуативно-обусловленной аббревиации можно считать примером того, как внешние условия коммуникации заставляют человека приспособлять письменную форму стандартного языка, создавая тем самым вариант языка для специальных целей. В более широком плане это пример воздействия общества на язык. Ситуативно-обусловленная аббревиация как результат этого воздействия и как закономерный способ компрессии языкового продукта занимает особое место в обеспечении акта коммуникации, поскольку характеризуется одной важной особенностью: понимание таких аббревиатур возможно только при условии знания конкретной обстановки, т. е. требует экстралингвистических сведений: «увеличение промпроизводства» (*газетн. промышленного производства*).

Аббревиация, как закономерное явление китайского языка, наиболее ярко проявляется в условиях отдельных подязыков, языков для специальных целей, и представляет собой определенную трудность для адекватного понимания текстов. Основная причина появления аббревиации заключается в возрастании влияния технологической сферы общества на развитие и изменение языка, что приводит к постоянному появлению новых компрессированных языковых единиц. Лишь небольшая часть из них находит отражение в словарях и справочниках, пополняя лексический состав языка. Однако большинство новых аббревизированных образований не выходит за рамки породивших их языковых подсистем. Они либо продолжают функционировать в составе отдельных микроподязыков, постоянно изменяясь и часто приобретая новые семантические значения, либо постепенно отмирают, выполнив свое коммуникативное предназначение и уступив дорогу новым языковым образованиям. Такое явление можно было бы назвать ведомственной аббревиацией, имея в виду, что редуцированный (сжатый) эквивалент единицы языка, созданный в пределах определенного ведомства или коллектива, существует только в пределах данного коллектива.

Так, в материалах исследования выявлено несколько десятков ведомственных аббревиатур, которые характерны только для данного микроподязыка. Срок жизни значительной части таких аббревиатур колеблется от 1 — 2 месяцев до 2 — 3 лет. Для введения аббревиатур в машинный словарь требуется, таким образом, определить срок их возможного существования. Включение в словарь аббревиатуры, которая уже закончила свое функционирование, может лишь ввести переводчика в заблуждение.

В целом анализ показывает, что исследуемый микроподязык, являясь подсистемой языка аналитического типа, обладает крайне бедно представленной морфологией. Основными средствами выражения синтаксических отношений между словами являются фиксированный порядок слов и служебные слова. Классическая формула структуры китайского предложения П — С — Д (подлежащее — сказуемое — дополнение) практически не нарушается, за исключением нескольких отклонений, имеющих строго обусловленный характер.

В процессе анализа установлено, что в целом на синтаксическом уровне исследуемый микроподязык характеризуется минимальным наличием сложных грамматических структур. Превалирующее положение занимают неполносоставные предложения клишированного характера, которые считаются нормой для данного микроподязыка.

Их широкое применение позволяет реализовать принцип экономии языковых средств одновременно с повышением скорости передачи информации и достижением практически стопроцентного уровня понимания без каких-либо искажений.

Следующим этапом построения функционально-коммуникативной модели является **информационный анализ**. Он позволяет опреде-

лить некоторые основные ситуации, которые описываются средствами данного микроподъязыка. Сначала создается список так называемых макроситуаций. В данном конкретном случае их установлено десять. Мы не будем перечислять их все, назовем лишь некоторые, чтобы дать общее представление о них.

1. Чрезвычайные происшествия.
2. Нарушения существующего порядка и правил судоходства.
3. Проведение ремонтных и прочих работ ...
6. Планирование грузопассажирских перевозок. ...
9. Прибытие делегаций и отдельных лиц.
10. Деятельность сопредельной стороны.

Далее каждая макроситуация может делиться на определенное число микроситуаций; например, макроситуация: «Чрезвычайные происшествия» может включать:

- столкновения судов;
- гибель людей и судоходных средств;
- несанкционированное пересечение линий границы;
- нарушение правил пограничного режима и т. п.

Для описания каждой макро- или микроситуации используется определенный набор устойчивых фраз, словосочетаний, отдельных слов (фрагментов текста), роль и значение которых заранее определены и строго фиксированы.

Лингвостатистическая модель микроподъязыка строится для того, чтобы:

- выработать методику отбора документов и определить необходимый размер выборки материала;
- выявить частотные характеристики;
- составить частотный словник статистически устойчивых элементов текста.

Статистические данные позволяют оптимизировать построение алгоритма автоматического анализа текста и в первую очередь определить наиболее важные и несущие основную информацию части текста.

Глава 8

МОДЕЛИ ПЕРЕВОДА И ВОЗМОЖНОСТИ ИХ ФОРМАЛЬНОЙ РЕАЛИЗАЦИИ

Комбинаторная и квантитативная лингвистика

Применение математических методов к изучению языка имеет свою историю, которая, однако, моложе истории языкознания как такового. Статистика, о которой мы упомянули в предыдущем пара-

графе, является лишь одной частью математических методов, применяемых к изучению языка. В предыдущих главах мы упоминали о том, что достаточно убедительные математические школы ставили задачу описать язык, его существование и действие с помощью аппарата и положений математической логики. Можно сказать, что в применении математики к решению лингвистических задач существуют два основных направления, которые можно назвать соответственно комбинаторной и квантитативной лингвистиками.

Выбор математического аппарата в лингвистических исследованиях зависит от того, как определяются основные задачи языкознания. Некоторые математики и лингвисты считают, что основной задачей лингвистики должно быть изучение грамматики, порождающей текст. При этом грамматика понимается как конечное множество детерминированных правил, в том числе неграмматических, а язык рассматривается как бесконечное число регулярных цепочек слов, порождаемых этой грамматикой. В такой постановке вопроса экспликация лингвистических явлений должна опираться на такие разделы неколичественной математики, как теория множеств, математическая логика, теория рекурсивных функций, теория алгоритмов и т. д. На основе применения этого аппарата сложилось направление так называемой **комбинаторной лингвистики** [Головин 1966].

Математическое языкознание не может лимитироваться только таким подходом, поскольку такое ограничение затрудняет преобразование нечетких лингвистических множеств, элементы которых имеют вероятностные веса принадлежности, в четкие множества искусственных языков. Между тем такое преобразование лежит в основе всех видов машинной переработки текстов. Кроме того, при таком чисто комбинаторном подходе вне сферы применения математических методов остается акустико-физиологическая и психолингвистическая проблематика речеобразования, а также стилистика и история языка, при изучении которых применяются не только и не столько комбинаторные, сколько количественные измерения.

Полезность и важность применения квантитативного подхода к изучению лингвистических объектов подчеркивали многие выдающиеся лингвисты прошлого и настоящего. «Нужно чаще применять в языкознании количественное, математическое мышление», — указывал Бодуэн де Куртенэ. Академик В. В. Виноградов, говоря о частотности употребления разных типов слов в разных стилях книжной и разговорной речи, отмечал, что «точные измерения в этой области помогли бы установить структурно-грамматические, а отчасти и семантические различия между стилями». «Частотность принадлежит функциональной стороне языковой системы... Учет частотности любого языкового явления — полезный прием при анализе» (В. Н. Ярцева) [цит. по: Головин 1966].

Квантитативные методы не в состоянии решать любые проблемы анализа языковых явлений. Такой подход способен охватить лишь

определенный аспект языка и речи. Но это — существенный аспект языка, отражающий ряд важных сторон речевой деятельности, который невозможно обнаружить чисто качественным подходом.

Всякая количественная характеристика лингвистических явлений предполагает их качественную характеристику. В то же время качественная определенность лингвистического объекта, в свою очередь, существенным образом зависит от количества элементов, его образующих, от частоты употребления или от силы взаимодействия (корреляции) элементов. Можно констатировать наличие тесной взаимосвязи качественных и количественных характеристик языка: совместное их рассмотрение открывает широкие лингвистические возможности исследования языковых процессов и явлений. Такой анализ допускает, например, прогнозирование качества со стороны количества и обратно. Открывается возможность изучения закономерностей корреляции количественных характеристик с качественными факторами (например, связь между частотностью и морфологическим строением слова). Во многих случаях количественные характеристики могут служить сигналом, направляющим внимание исследователя на некоторые скрытые от простого наблюдения качественные особенности и закономерности индивидуальных или функциональных стилей.

Можно заметить, что количественные исследования в настоящее время служат основанием для становления и развития такой отрасли лингвистической науки, как лингвистическая синергетика [Пиотровский 2006]. Р. Г. Пиотровский, основоположник творческого применения статистических методов в лингвистике, пишет, что с помощью анализа больших текстовых массивов были выявлены следующие модельные зависимости, которые можно рассматривать в качестве внешних симптомов работы синергетических механизмов:

- зависимость, описывающая так называемый диахронический скачок, т. е. интенсивный рост употребительности некоторых лексических единиц или грамматических форм при переходе языка из одного типологического состояния в другое, так называемый закон Пиотровских;

- зависимость между уровнем энтропии текста и его принадлежностью к тому или иному стилю речи, а также между величиной энтропии в норме языка и уровнем владения языком учащегося;

- обратная зависимость между частотой лингвистической единицы и ее номером в частотном словаре, так называемый закон Ципфа-Мандельброта и его модификации;

- прямая зависимость между средней длиной словоформы, измеренной в слогах или морфемах, и энтропией распределений ее слоговых и морфемных длин;

- обратная зависимость между длиной текстовой цепочки (слова, словоформы, словосочетания и т. п.) и средней длиной составляющих ее слогов, морфем или словоформ;

– прямая зависимость между частотой словоформы и ее многозначностью, а также употребительностью в разных контекстах.

Лингвистические основы исследования

В качестве основы лингвистического подхода к количественным характеристикам языка можно выделить принцип разделения языка и речи. Возможность и необходимость разграничения этих двух сторон языковой деятельности — языка и речи — основаны на том очевидном факте, что в этой деятельности можно различить два взаимосвязанных, но отделяемых друг от друга компонента: средство (орудие) и его применение.

Учитывая специфику количественно-системного исследования языка вообще и лексики в частности, можно представить комплекс «язык-речь» в виде перекрещивающихся двух главных осей: оси с противопоставлением потенции-реализации и оси с противопоставлением динамики-статики [Тулдава 1987] (рис. 8).

В соотношении потенции-реализации содержится элемент деления на уровни: потенция — высший уровень, реализация — низший. Следует отметить, что в онтологическом плане дело обстоит как раз наоборот: речь предшествует языку; как конкретная реализация языка, она является единственным прямо наблюдаемым объектом языкознания.

В количественной лингвистике противопоставление языка и речи имеет прямой практический смысл.

Реализация языка может рассматриваться как актуализация или выбор одного из возможных параметров в данной ситуации. Например, выбор подходящего слова из данного семантического поля. В этом смысле можно противопоставить словарь (лексикон) и текст.

С понятиями потенции и реализации как характеристиками языка и речи связывают понятия вероятности и частотности. «Язык вероятностен, речь частотна» (Б. Н. Головин).

Характеристики динамики и статики часто используются при разграничении речи как процесса и речи как результата этого процесса. Однако уровень языка рассматривается чисто в статистическом плане как «инвентарь языковых средств и набор правил». Разницу между сферами динамики и статики видят еще и в том, что динамика (механизм и процесс) связывается с деятельностью мозга, а статика (язык как предмет) находится вне человека.



Рис. 8. Язык и речь

Признаки	Динамика	Статика
Потенция	Языковая компетенция	Языковая схема (язык)
Реализация	Речевой процесс	Речевой продукт

Рис. 9. Аспекты речевой деятельности

В итоге в системе речевой деятельности можно различать четыре основных аспекта (подсистемы): на языковом уровне — языковую компетенцию и языковую схему, на речевом уровне — речевой процесс (акт) и речевой продукт (текст).

Названные выше элементы выявляются вследствие того, что парные признаки — потенция-реализация и динамика-статика — накладываются друг на друга, перекрещиваясь, в результате чего образуется структура, аналогичная четырехпольной таблице сопряженности альтернативных признаков. На рис. 9 показаны эти аспекты речевой деятельности.

Рассмотрим содержание данных аспектов.

Языковая схема. Под ней понимается система языковых элементов и отношений между ними. Последние могут относиться как к синтагматике, так и к парадигматике. Языковую схему можно рассматривать как статическую систему данного языка в целом, т. е. как суммарную совокупность лингвистических элементов и отношений между ними. Языковая схема отличается устойчивостью в рамках данного языка. В ней выделяется, например, общая лексика, присущая всем подъязыкам, она характеризуется стабильными связями и закономерностями на общеязыковом уровне. Выработавшееся в общественной практике типичное и общепринятое языковое употребление, регулярно повторяющееся в определенной сфере коммуникации и отражающееся в виде некоторого устойчивого «ядра» подсистемы, называется *нормой* для данного языка. В количественной лингвистике норму можно определить как наиболее вероятный состав элементов и наиболее вероятные отношения между элементами.

Языковая компетенция является следствием отражения сознанием людей языковой схемы, т. е. это набор элементов и системных отношений между ними, сходных (в плане отражения) с элементами и отношениями языковой схемы, плюс особый динамический компонент, необходимый для приведения языка в действие. Без обращения к соседним наукам лингвистика не может объяснить действительную природу и внутренние закономерности речевой деятельности в целом. Точно так же количественные закономерности речи (универсальные особенности статистической структуры текста, наличие устойчивых распределений и регулярности в корреляции с внешней средой и пр.) могут быть объяснены только при условии привлечения к анализу экстралингвистических представлений и кри-

териев. В частности, должны быть привлечены данные психологии, физиологии нервных процессов и т. п.

Речевой процесс (акт) является реализацией потенций языка, т. е. самым действием механизма порождения речи. Отношение речевого процесса к языковой компетенции можно рассматривать как отношение управляемой системы к управляющей. Речевой процесс имеет своим непосредственным результатом линейную последовательность речевых единиц. В зависимости от исследовательской задачи единицами речи можно считать звуки, слова, фразы и т. п. На лексическом уровне за основную единицу следует принять слово, соответствующее минимальной смысловой единице в составе высказывания. По формулировке А. А. Леонтьева, «сам процесс порождения высказывания представляет собой сложный, многоступенчатый речемыслительный акт, состоящий из этапов превращения исходного замысла через внутреннюю речь в схему речевого высказывания и включения его в фонематические, лексико-семантические и логико-грамматические коды языка».

Речевой продукт. Результатом процесса порождения речи является речевой продукт, т. е. текст, который в самом общем виде понимается как определенным образом (обычно письменно) зафиксированный отрезок речевого континуума. По отношению к речевому процессу текст характеризуется тем, что в нем снята процессуальная форма речемыслительной деятельности, хотя он и сохраняет связь с этой деятельностью. Будучи единственным прямо наблюдаемым объектом лингвистического исследования, текст может при соответствующем подходе служить и моделью для изучения динамических сторон речи. В этом смысле можно расширить понимание текста, включив в него особым образом отраженные компоненты речевого процесса — темы и ситуации. При анализе текста можно также выявить в нем рефлексии языковой компетенции, воплощенные в виде определенных стилистических и прагматических черт.

В целом текст и соответствующий словарь могут рассматриваться как вероятностные системы. Эмпирически это проявляется, с одной



Рис. 10. Модель системы речевой деятельности

стороны, в наличии стабильных распределений единиц, в устойчивых корреляциях внутрисистемных и межсистемных объединений, в образовании «ядра» в замкнутых группах, а с другой — в явлениях периферии, размытости границ и различного рода флуктуациях.

Таким образом, на основе выделения перекрещивающихся осей потенции-реализации и динамики-статики общую систему речевой деятельности можно разделить на отдельные части, которые рассматриваются как основные подсистемы в плане функционирования системы. При этом устанавливается внутренняя структура, взаимосвязь между отдельными подсистемами.

Модель системы речевой деятельности, представленная на рис. 10, дает некоторое общее представление о структуре, связях и взаимоотношениях компонентов языка и речи, описанных выше.

Методика количественного исследования

Квантитативно-системный анализ характеризуется выбором и комбинацией статистических методов, специально приспособленных к лингвистическим исследованиям, но в то же время характерных для математической статистики как специальной дисциплины в рамках точных наук.

Предпосылкой применения квантитативных методов является **квантификация** исследуемого материала. В широком смысле она означает количественное представление качественных явлений, т. е. такую процедуру, при которой исследуемым явлениям, имеющим принципиально качественную природу, например лингвистическим объектам, приписываются количественные оценки, вследствие чего эти объекты и явления могут быть изучены как количественные. Как правило, квантификация включает предварительное **квантование** объекта, т. е. его приведение в удобную для измерения форму, выделение единиц учета на разных уровнях исследования и последующее измерение.

Квантование — предварительный этап квантификации. Например, непрерывные переменные величины могут быть приведены к дискретному виду посредством разбиения области их изменения на совокупность непересекающихся интервалов. Любые процедуры квантования должны быть основаны на содержательных соображениях и однозначным образом интерпретироваться.

Измерение — процедура, приписывающая по заранее фиксированным правилам числовые значения тем или иным наблюдаемым объектам. Своим логическим основанием оно имеет категории свойства и отношения. Говоря о количественном представлении объекта на основе какого-либо свойства, мы в сущности приписываем количественное значение не самому объекту, а его свойству или признаку.

Кроме первичных, прямых измерений выделяются еще и производные (косвенные). Измерения называются первичными, если они не основаны на каких-либо предварительных измерениях: в противном случае они именуется производными. В качестве классического примера производного измерения можно привести установление скорости движения — скорость определяется как отношение длины пройденного пути и времени, за которое этот путь был пройден. В количественной лингвистике производным измерением считается, например, «индекс разнообразия» (отношение объема словаря к объему текста, коррелятивная функция как отношение условной вероятности к независимой вероятности и многие другие функции или отношения частот).

Проиллюстрировать эти положения можно ссылкой на проведенные голландскими учеными в 2007 г. исследования особенностей и закономерностей, отличающих мужскую и женскую устную речь в спонтанном виде и в радио- и телепередачах [Wiggers и др. 2007]. Были обследованы записи разных видов речи и затем статистически сравнены полученные образцы по основным параметрам сходства и расхождения. Авторы исследования не исходили из каких-либо предварительных гипотез, а базировались на изучении массивов записанных текстов. Они пришли к выводу, что мужская и женская речь различаются по ряду параметров, а именно: женщины чаще используют междометия в обоих видах речи, как спонтанной, так и подготовленной (радио и телевидение); мужчины употребляют больше сложных слов. Женщины чаще говорят о других людях, мужчины — об объектах. Отмечены различия в употреблении различных слов в разных объемах текстов, так называемый коэффициент разнообразия (type-token ratio), а также получены другие интересные данные путем применения статистических методов к обследованию материала в рамках соответствующего подхода (data mining approach).

Измерения могут производиться по разным уровням или шкалам: номинальной, порядковой, интервальной и пропорциональной. Каждая из них характеризуется соответствующей числовой системой и возможными (допустимыми) эмпирическими операциями.

Моделирование с помощью распределений

Распределение, как обобщающее, интегральное понятие, является важнейшей структурной характеристикой именно вероятностных систем. Поэтому моделирование с помощью распределений, т. е. построение моделей, является методом системного изучения внутреннего строения объекта или его поведения. Распределение отражает не только наличие внутренней упорядоченности в системе, но и взаимодействие между элементами системы и представляет собой, та-

ким образом, синтез внутренней расчлененности и общности в строении системы.

Само понятие «распределение» можно понимать в широком смысле как упорядоченную совокупность количественно выраженных значений, т.е. результатов измерений объекта, обычно с указанием значимости (частоты, вероятности, ранга) этих значений в данной совокупности. В более узком смысле вероятностное распределение величины определяется как перечисление значений случайной величины и их вероятностей.

За основу при изучении распределений целесообразно взять методологический принцип разделения языка и речи, сфер статики и динамики в рамках общей системы речевой деятельности.

Учитывая выделение в системе речевой деятельности уровней языка и речи, распределения могут быть разделены на **теоретические** (языковые) и **эмпирические** (речевые).

Соответственно распределения можно разделить на **статические** и **динамические**. С точки зрения внешней формы они различаться не будут, однако приобретут разный содержательный смысл. Статические распределения, называемые также синхронными, выражают преимущественно синхронический и парадигматический аспекты анализа лингвистических явлений. Сюда относится, например, ранговое распределение слов, или распределение частот слов в тексте (частотный спектр лексики). В то же время в качестве статических распределений можно рассматривать разбиения некоторых совокупностей, например лексико-семантических полей, на отдельные группы (классы, кластеры), упорядоченные с точки зрения их вероятностно-частотных свойств.

Динамические распределения отличаются от статических тем, что они выражают либо процессуальность, связанную с явлениями порождения речи (в синхронии), либо изменения в развитии языка (в диахронии). Во всех случаях явно или неявно учитывается момент времени как изменяющийся фактор в составе распределений, и поэтому динамические распределения часто именуют также диахронными в широком смысле слова.

Таким образом, кроме разделения распределений на теоретические (языковые) и эмпирические (речевые) выделяются характерные для лингвистики статические (синхронные) и динамические (диахронные) распределения. Наряду с этими основными типами лингвистических распределений можно различать еще ряд подтипов, или разновидностей, распределений.

Выборка. Необходимой предпосылкой применения математической статистики является образование репрезентативной выборки на основе принципа случайности, в соответствии с которым всем единицам генеральной совокупности должна быть предоставлена возможность попасть в выборку. Выполнимость принципа случайности связана с характером генеральной совокупности, с ее качеством и объемом.

По характеру генеральные совокупности делятся на **статистически однородные** (однородные) и **статистически неоднородные** (разнородные). В однородной совокупности на разных ее участках действует один и тот же комплекс систематических факторов (или один систематический фактор). Именно это постоянство обеспечивает близость статистических показателей различных выборок, взятых из одной и той же генеральной совокупности. Количественное расхождение между различными выборками, взятыми из однородной совокупности, создается случайными факторами, которые то усиливают, то ослабляют действие систематических факторов. В неоднородной совокупности комплекс систематических факторов непостоянен, на разных ее участках он так или иначе изменяется. Одни факторы могут прекращать свое действие, другие — появляться. Статистические показатели различных выборок неоднородной совокупности значительно отличаются друг от друга.

По объему генеральные совокупности делятся на **конечные** и **бесконечные**. Если совокупность содержит ограниченное число единиц, она называется конечной. Если число единиц безгранично, то она является бесконечной. В лингвистике целесообразно придать этим терминам некоторый иной смысл. Конечной совокупностью называют практически обозримый отрезок речи, относительно небольшой, четко очерченный, целиком доступный для статистической обработки, — отдельная книга, список сочинений одного автора, сумма статей из журнала на определенную тему и пр. Бесконечной совокупностью можно назвать практически необозримые отрезки речи, например все тексты данного языка или его стилистической разновидности, не только в каких-то временных рамках, но и за небольшой период времени.

Связь между характером генеральной совокупности и приложимостью статистических методов. Если генеральная совокупность конечна, то статистику можно применять без ограничений, поскольку в этом случае существует возможность сделать репрезентативную выборку независимо от того, однородна данная генеральная совокупность или нет. Наиболее уместным способом является механический отбор. При этом вся генеральная совокупность разбивается на равные части, из каждой берется одна единица (индивидуальный механический отбор) или целая серия (серийный отбор). Механический отбор легко организовать, он хорошо улавливает изменения в составе генеральной совокупности.

Если генеральная совокупность однородна или не слишком разнородна, то предпочтением заслуживает серийный отбор. Он требует меньшей затраты труда и облегчает качественный анализ. Если же генеральная совокупность сильно неоднородна, то предпочтительнее индивидуальный отбор. Обследование с помощью выборок не дает абсолютно достоверного результата: характеристики генеральной совокупности, полученные на основе выборки, должны сопровождать-

ся указанием на возможную величину ошибки репрезентативности и на определенную степень вероятности того, что эта ошибка не превысит заданной величины. Математическая статистика позволяет рассчитать объем выборки, необходимый для обеспечения очень высокого, практически вполне достаточного уровня достоверности: от 5 до 20 % генеральной совокупности. При очень большом объеме генеральной совокупности качественный анализ всех ее единиц практически неосуществим, поэтому необходима выборка. Сплошное обследование возможно только тогда, когда генеральная совокупность практически мала и доступна.

Рассмотрим бесконечную генеральную совокупность. В этом случае возможности статистических методов существенно ограничены. Для исчисления ее обобщающих характеристик вероятностно-статистические методы применимы только тогда, когда бесконечная совокупность однородна. В противном случае эти методы способны только зафиксировать сам факт неоднородности. Бесконечная совокупность не может быть положена целиком в основу выборки, поэтому выборка производится из какого-либо фрагмента генеральной совокупности, который, как бы он ни был велик, составляет лишь часть совокупности. Например, при изучении научного стиля определенного языка в определенную эпоху невозможно произвести отбор из всех научных книг и журналов всех научных дисциплин. Приходится ограничиваться сравнительно небольшим списком текстов, в несколько сотен или тысяч страниц.

Выборка из какого-то фрагмента бесконечной совокупности непосредственно представляет только данный фрагмент, и вычисленные характеристики относятся только к нему. Распространение характеристик фрагмента совокупности на всю совокупность правомерно лишь в том случае, если их качества совпадают, т.е. если вся генеральная совокупность однородна.

Многочисленные исследования текстов с помощью методов математической статистики показали, что язык как совокупность функциональных стилей не имеет ни одного количественного признака, относительно которого он был бы однороден. Каждый функциональный стиль тоже, как правило, неоднороден. Б. Н. Головин пишет, что «опыты показывают, что главные функциональные стили имеют внутреннюю дифференциацию на жанровые, тематические и даже индивидуальные разновидности. Так, разновидность публицистического стиля, соотношенная с жанром заметок и сообщений, может по ряду признаков существенно отличаться от разновидности, соотношенной с жанром передовой статьи... Однородность в том или ином отношении нередко обнаруживают отдельные разновидности функциональных стилей. Так, по данным Якубайтис и Стурие, в латышской художественной литературе стиль прозы, стиль драматургических произведений и стиль поэзии однородны по пяти признакам: по количеству разных грамматических форм, по количеству глаголов

и существительных, по количеству существительных в винительном и дательном падежах» [цит. по *Тулдава* 1987].

Исследования по изучению качества бесконечных генеральных совокупностей с помощью статистических методов важны не только потому, что они выделяют те аспекты речевой деятельности, при описании которых статистика играет важную роль. Такие исследования содействуют более четкой и надежной дифференциации различных типов речи и дают представление о характере действующих в них качественных факторов.

Интерпретация статистических распределений

Лингвистические распределения являются чисто формальными моделями, и их надо рассматривать как интерпретируемые знаковые системы, т.е. они требуют качественной интерпретации. В рамках таких интерпретаций выявляются основные свойства и отношения внутри систем, а также те закономерности, которые определяют характер лингвистического явления.

Интерпретацию лингвистических распределений можно представить как многоуровневый процесс качественно-количественного анализа и синтеза. Задача состоит в том, чтобы раскрыть сущность явления, однако здесь могут быть разные пути к достижению цели. Различаются индуктивный и дедуктивный подходы. Высшей формой интерпретации является научное объяснение, которое должно показать, что данный научный факт является проявлением определенного закона или что объясняемый закон является следствием более общей теории. В зависимости от условий и задач исследования и характера материала при исследовании лингвистических распределений Ю.Тулдава выделяет три типа интерпретации: структурно-функциональную, прагматическую (стилистическую) и генетическую (причинную).

Структурно-функциональная интерпретация. В рамках этой интерпретации есть два уровня: индуктивный и дедуктивный. На первом практикуется графико-геометрический метод изображения статистических данных. Примером может служить закон Ципфа-Мандельброта, отражающий зависимость между рангом слова в частотном словаре и вероятностью (частотой) его употребления. Чаще всего этот закон изображается в системе координат в виде некоторой кривой линии. В рамках этого подхода возможно применять формулы. Они допускают проведение различных математических операций, в результате которых могут появиться новые данные или новые аспекты закономерностей, выявленных статистически. В качестве аналитических записей распределений в квантитативной лингвистике обычно употребляют различные функции или дифференциальные уравнения.

На дедуктивном, гипотетико-теоретическом уровне используются модели вероятностных систем, когда те или иные функции модели выводятся теоретически, на основе некоторых постулатов большей или меньшей степени общности. Интерпретация таких моделей содержит исходные постулаты и возможные выводы, сделанные на основе анализа конкретного материала. Например, в лингвистике известны попытки вывести закон Ципфа из общих закономерностей вероятностного подхода или по аналогии с термодинамикой.

В квантитативной лингвистике возможен также гипотетико-эмпирический (промежуточный) подход, при котором исходная модель получается на основе некоторых относительно самостоятельных теоретических схем, т. е. выдвигается некоторая гипотеза, а затем с помощью эмпирических данных определяется конкретная форма распределения лингвистических единиц в соответствии с требованиями предложенной модели.

Прагматическая (стилистическая) интерпретация по существу является объяснением, которое основывается на связи между формальными (квантитативными) показателями и некоторыми прагматическими, т. е. оценочными, характеристиками изучаемых явлений. При этом могут быть использованы такие содержательные, стилистические понятия, как «богатство» или «разнообразие» словаря, «целостность» (художественная завершенность) текста и пр. В таком случае характеризуются некоторые свойства объектов, при этом реально не наблюдаемые и прямо не измеряемые. При анализе обращаются к внешним сторонам наблюдаемых объектов (словарей, текстов) и пытаются добраться до сути явлений.

Необходимо иметь в виду, что связь количественных индикаторов с латентными переменными (свойствами, характеристиками) имеет вероятностный характер. Это означает, что интерпретация может дать положительный эффект только на представительном материале и что формальные индикаторы могут лишь косвенно отражать содержательные свойства явлений.

Генетическая (причинная) интерпретация. При таком подходе пытаются объяснить изучаемое явление, указывая на его происхождение. Тем самым прямо или косвенно указывается причина его существования. Вероятностный подход определяет вероятностную концепцию причинности, согласно которой причинность может присутствовать в большей или меньшей степени, а не только быть или не быть. Признается также принцип множественности причин, который обусловлен множественностью связей данного явления с другими. В настоящее время многие лингвистические явления объясняются с точки зрения других наук, так или иначе связанных с лингвистикой: психолингвистики, психофизиологии, филологии, обществоведения, истории и др. В качестве примера можно привести названия некоторых диссертаций по лингвистическим специализа-

циям «Теория языка» и «Сравнительно-историческое, типологическое и сопоставительное языкознание», защищенных в 2009 г.: «Вербальная характеристика агрессии личности», «Языковые способы кодирования поспессивности», «Сопоставительный анализ лексики нравственных отношений в английском и русском языках», «Способы передачи сказовой манеры Н. С. Лескова с русского на английский язык», «Особенности перевода эллиптических предложений на материале переводов драматургии А. Н. Островского», «Роль компрессии в лексической номинации» и т. п. Практически в каждой из этих диссертаций присутствует некоторая статистическая основа исследования, развитая далее в рамках примыкающих к лингвистике других наук.

Гиперболическое распределение частот слов, известное под названием закона Ципфа, связывают иногда с особенностями психики человека, его потребностью в общении, с одной стороны, и со стремлением свести к минимуму свои умственные и физические усилия, с другой. Этот принцип «наименьшего усилия» основывается на взаимодействии двух противоборствующих тенденций в психике (подсознании) человека. Представители психофизиологии некоторые типы лингвистических распределений как моделей речевой деятельности связывают со структурными особенностями мозга, в частности с пространственно-временной организацией периодических (циклических) процессов в мозгу. Порождение речи рассматривается иногда как результат логико-лингвистического синтеза (см., например, работу Е. С. Андреевой [Андреева 2001]). В ней представлены: эскиз системной сетевой логики на основе принципов диалектики Гегеля; понятие как системная логическая форма; «логический, или предикативный, треугольник» как элементарный структурный модуль системы понятий, ячейка понятийной сети, механизм порождения слова и элементарная система; процесс развития триединой системы «мышление-язык-знание» как движение от абстрактной сети к идеальной конкретной точке и т. п.

В филогенетическом плане лингвостатистические распределения объясняются эволюционным развитием человеческого языка, который формировался на протяжении тысячелетий в результате приспособления к внешнему миру [Гринев-Гриневиц и др. 2005].

Такой эволюционный процесс можно уподобить органической эволюции на базе естественного отбора. Системное свойство основных лингвистических распределений, а именно иерархичность, можно объяснить адаптацией к увеличению числа элементов (при порождении речи), учитывая, что иерархическая структура минимизирует число связей. Устойчивость некоторых лингвистических распределений (сохранение общей формы) может указать на стремление к равновесию, оптимальности и целесообразности самоорганизующейся сложной системы — языка.

Модели «человеческого» («ручного») перевода

Как было уже сказано выше, ни «человеческий», ни «ручной» перевод как термины не очень подходят для описания такого вида перевода, который, в отличие от машинного, осуществляется человеком. В этом смысле английское слово *human* имеет более широкий спектр значений и удовлетворяет требованию точности передачи. Слово «человеческий» в русском языке тоже многозначно, но его коннотации несколько иные, а термин «ручной» вообще не очень удачен потому, что перевод производится не руками, а головой. Тем не менее более точного термина для обозначения данного вида перевода пока не найдено.

В книге Л. Л. Нелюбина [*Нелюбин 2007*] подробно описывается история переводческого дела с давних времен, практически с момента зарождения человеческих цивилизаций — египетской, древнегреческой, латинской. Особую роль в развитии и становлении переводческого дела сыграли мировые религии — язычество, христианство, буддизм. Перевод религиозных текстов выдвигал высокие требования к качеству. Нет необходимости говорить о том, что предварительное понимание таких текстов с целью их точного перевода требовало высокой грамотности и культурно-религиозной компетенции переводчиков. Впервые на необходимость создания теории перевода обратил внимание Марк Туллий Цицерон, известный деятель Древнего Рима. История перевода и переводоведения ярко показывает неразрывную связь перевода с тем, что мы называем человеческим интеллектом. Перевод — языковая деятельность, а язык и естественный интеллект представляют собой единое целое [*Рябцева 2005*].

Кратко рассмотрим модели перевода, перечисленные и детально описанные Л. Л. Нелюбиным в вышеназванной работе.

Теория закономерных соответствий, сформулированная Я. И. Рецкером, учитывает то обстоятельство, что некоторые приемы логико-семантического порядка в процессе перевода повторяются. Так, для перевода отдельных слов и словосочетаний могут существовать устойчивые эквиваленты в языке перевода. Для других типов лексических единиц могут применяться определенные контекстные ситуации при выборе вариантов перевода — такие соответствия называются вариантными или контекстуальными. Наконец, в сложных случаях могут потребоваться сложные трансформации на языковом уровне, связанные с семантикой и закономерностями выходного языка.

Информационная теория перевода разработана Р. К. Миньяром-Белоручевым. По этой теории, в процессе перевода происходит не поиск и осуществление трансформаций, а поиск и передача информации. Информационная модель перевода основана на учете информационного запаса пяти степеней получения информации: уникальной или ключевой, дополнительной, уточняющей, повторной и нулевой. Слово «информация» в этой теории имеет несколько зна-

чений. В переводе должна сохраняться не вся информация, а только та, которая предназначена для передачи. Она называется «инвариантом перевода».

Семантико-семиотическая модель перевода. Автором ее Л. Л. Нелюбин называет Л. С. Бархударова, который развил теорию закономерных соответствий Я. И. Рецкера и построил всеобъемлющую типологию корреляций знаков и структур исходного и переводящего языков. Согласно этой концепции, в общем содержании текста выделяются отдельные его компоненты — денотативный, сигнификативный, внутрилингвистический и прагматический. Впервые было показано, что в переводном тексте сохраняется не все содержание оригинала, а лишь те или иные его компоненты.

Ситуативная (денотативная) модель перевода несколько выходит за рамки межъязыковых преобразований и предполагает обращение к действительности. Она исходит из того, что любой денотат (предмет, признак, действие, т. е. элемент окружающей действительности), любая ситуация могут быть описаны по-разному. При этом в каждом языке есть свои устоявшиеся способы их описания, которые надо находить не путем преобразования отдельных единиц текста, а через описываемый денотат, через описываемую ситуацию. В рамках этой теории переводчик действует как высококвалифицированный лингвист, владеющий теорией и практикой двух языков. Эта теория активно разрабатывается в трудах В. Г. Гака.

Теория уровней эквивалентности объединяет теорию закономерных соответствий, информационную теорию, семантико-семиотическую и ситуативно-денотативную модели в рамках некоторой единой теории. Такую попытку предпринял В. Н. Комиссаров. В основе этой теории лежит представление о том, что степень реальной смысловой близости между оригиналом и переводом является величиной переменной. В содержании текста выделяется ряд последовательных уровней, отличающихся по характеру информации, а отношения эквивалентности устанавливаются между аналогичными уровнями содержания текстов исходного и переводящего языков.

В. Н. Комиссаров выделяет следующие типы эквивалентности отношений между текстами оригинала и перевода: эквивалентность на уровне цели коммуникации, характеризующаяся наименьшей общностью содержания оригинала и перевода; эквивалентность на уровне описания ситуации; эквивалентность на уровне способов описания ситуации, при которой сохраняются понятия, с помощью которых была описана ситуация в исходном тексте; эквивалентность на уровне структурной организации высказывания, при котором к вышесказанному добавляется инвариантность синтаксических структур оригинала и перевода; эквивалентность на уровне семантики словесных знаков — наименьшая степень смысловой общности, которая вообще может не существовать между оригиналом и переводом.

Иллюстрируя проблему эквивалентности [Комиссаров 2004], Комиссаров приводит такой пример. Английское предложение *The student is reading a book* можно перевести на русский язык как *Студент читает книгу*. Вроде бы все правильно и другой перевод трудно себе представить. Однако очевидно, что полного тождества здесь нет. В английском оригинале содержится указание, что речь идет о каком-то определенном лице, известном собеседникам, и, напротив, о какой-то неопределенной книге. В переводе такая информация отсутствует, как и информация о том, что действие происходит в момент речи, а не повторяется регулярно. В то же время, русский язык вынуждает нас сообщить, что читающий — это лицо мужского пола, хотя в оригинале речь могла идти и о женщине. Все эти потери и добавления обычно не будут препятствовать переводу выполнять свою функцию, заключает В. Н. Комиссаров.

Однако ясно, что говорить о том, что такой перевод полностью эквивалентен, можно лишь с учетом предыдущих соображений.

Возвратимся к оценке модели В. Н. Комиссарова в работе Л. Л. Нелюбина. Он пишет, что В. Н. Комиссаров разработал универсальную типологию семантических состояний, в которых может пребывать исходный текст относительно текста перевода. Для каждого из этих состояний постулируется свой выбор лексико-семантических трансформаций. В переводческой модели В. Н. Комиссарова в наиболее полной мере реализован и использован когнитивный потенциал собственно лингвистического подхода к переводу как процессу межъязыковых преобразований.

Функционально-прагматическая (динамическая) модель перевода. Одним из основоположников коммуникативного подхода к переводу вообще и к исследованиям его функционально-психологических аспектов в частности является А. Д. Швейцер. Он дает описание процесса принятия переводчиком решений на основе той или иной конфигурации языковых и внеязыковых характеристик текста, который складывается из двух основных этапов: выработки стратегии перевода и определения конкретного переводческого воплощения этой стратегии (сюда относятся различные конкретные приемы — «переводческие трансформации», — составляющие технологию перевода). Первый этап является своего рода прагматической надстройкой в рамках собственно лингвистического подхода к переводу. Наличие этого этапа делает осмысленной и целенаправленной операционную структуру деятельности переводчика, как определяет это Л. Л. Нелюбин. Например, при выработке стратегии перевода предпочтение может быть отдано текстуально точному переводу, приближающемуся к буквальному или, напротив, смело отходящему от формальной структуры оригинала, приближающемуся к вольному. В выработку стратегии перевода, по А. Д. Швейцеру, входит также принятие решения относительно тех аспектов оригинала, которые должны быть отражены в первую очередь. Наконец, решение относительно

но стратегии перевода включает еще один выбор: скрупулезная передача экзотических деталей местного колорита эпохи либо отказ от этого в пользу глубинного проникновения в историческую специфику текста.

Прагматические факторы, влияющие на процесс перевода, следующие: коммуникативная интенция отправителя, установка на получателя, коммуникативная установка переводчика. Надо заметить, что эти же моменты входят в стратегию перевода в том виде, как ее определил Ю. В. Рождественский.

Инвариантным в процессе перевода в рамках рассматриваемой модели является смысл, формируемый контекстом и ситуацией общения. Именно речевой контекст и ситуация общения дают возможность нейтрализовать различия между нетождественными значениями или использовать разные значения для передачи одного и того же смысла.

Перевод в концепции А. Д. Швейцера детерминирован множеством языковых факторов, которые образуют ряд взаимосвязанных фильтров, определяющих его стратегию. К этим факторам относятся система и норма двух языков, две культуры, две коммуникативные ситуации — первичная и вторичная, предметная ситуация, функциональная характеристика исходного текста, норма перевода. Сюда можно было бы добавить различные языковые картины мира. Центральное место во всем этом принадлежит тексту, который является точкой приложения сил, исходящих от детерминант перевода, и в то же время сам характеризует первичную и вторичную коммуникативные ситуации.

Психолингвистическая теория перевода сформулирована А. Ф. Ширяевым в рамках коммуникативного подхода. Согласно этой теории, цель перевода заключается в создании продукта в виде реально озвученного или записанного высказывания или группы высказываний на определенном языке. Социальным заказом переводчику задаются основные параметры производимого продукта. Объектом деятельности является выраженный средствами исходного языка продукт речевой деятельности одного из разноязычных коммуникантов. Средства переводческой деятельности включают в себя средства не одного, а двух языков.

Способы осуществления переводческой деятельности подразделяются на три группы. Первая группа — слушание и говорение. Вторая группа — процесс решения переводческих задач — использование эквивалентов, установление вариантных и контекстуальных соответствий. Третья группа — говорение и письмо. Структурно переводческая деятельность состоит из последовательности взаимосвязанных и взаимообусловленных действий, каждое из которых имеет три фазы: ориентировка в условиях деятельности и выработка плана действий; реализация выбранного плана; сопоставление результата с намеченной целью. Продуктом переводческой деятельности является текст на переводящем языке.

Коммуникативно-функциональная теория перевода. Заметим, что в изложении Л. Л. Нелюбина термины «теория» и «модель» употребляются практически как синонимы, однако в некоторых случаях «модель» понимается как результат реализации теории, что вполне правомерно. Модель не может возникать без определенной теории, в то время как теория может и не приводить к созданию модели. Надо сказать, что в «человеческой» теории перевода многие термины определяются по-разному, такие, например, как «адекватность», «эквивалентность», «равноценность» и пр. [см. *Комиссаров 2004*].

Коммуникативно-функциональная теория базируется на разграничении и соотношении понятий языкового значения и речевого смысла. Суть действий переводчика определяется его ключевым положением в цепочке многократного перекодирования информации, составляющей основу процесса двуязычной коммуникации: «смысл-значение-смысл-значение-смысл». В процессе перевода необходим учет объективных и субъективных факторов, формирующих конкретную речевую ситуацию общения. По мнению З. Д. Львовской, автора этой модели, основные причины переводческих трансформаций носят лингвистический и экстралингвистический характер.

Герменевтическая модель перевода. Герменевтика — наука о толковании (понимании) текстов. В этой модели перевод подчиняется закону понимания: начинается с понимания и завершается им. Переводчик осуществляет повторное понимание того, что им уже понято, в расчете на иноязычного получателя текста перевода. Герменевтическая модель перевода связана с интерпретационной трактовкой сущности процесса перевода, понимаемого как процесс вторичного порождения текста.

Модель социальной регуляции переводческой деятельности относится к исследованию социального аспекта коммуникативного подхода к переводу. Перевод в этом аспекте представлен как процесс установления и поддержания целенаправленного контакта между личностями в структуре межъязыкового общения. Адекватность перевода понимается как степень соответствия избранных переводчиком вербальных и невербальных средств социальным условиям общения, в результате чего у иноязычного коммуниканта вызывается реакция, соответствующая коммуникативной установке отправителя.

Теория военного перевода. Представляется, что в данном случае Л. Л. Нелюбин исходит из несколько иной позиции, чем в предыдущей классификации моделей и теорий, когда за основу брался некоторый абстрактный подход. Здесь же имеет место своего рода конкретно-жанровая манифестация функционально-прагматической концепции перевода. Если принять такой подход к классификации теорий и моделей перевода, то тогда надо выделять перевод медицинский, сельскохозяйственный, машиностроительный, химический и т. п. Однако надо заметить, что военный перевод действительно обладает весьма специфическими особенностями, оправдывающими

выделение его в самостоятельный вид. В нем особенно возрастает роль ошибки. Если в других видах перевода ошибка может и не приводить к каким-то сложным последствиям, то в военном переводе она должна быть совершенно исключена.

Культурологическая концепция перевода. Рассмотрение культурологических особенностей, связанных с общественными институтами (см. социальные аспекты перевода), обусловлено тем, что человек как социальное существо не мыслит себя вне зависимости от общества. В соответствии с культурологической концепцией В. И. Хайруллина специфика культуры находит неоднозначное проявление в текстах, подлежащих переводу. При этом имеется в виду особая языковая картина мира, т. е. своеобразное представление действительности в языке. К этому относятся опосредованные культурные факторы, а именно специфические когнитивно-семантические структуры, представляющие своеобразие структур мышления представителей разных структур. Такой план рассмотрения называется «культура в языке».

Второй план — «культура, описываемая языком». Культура понимается как совокупность материальных и духовных особенностей народа, которые определяют специфику организации его мышления, общественной и повседневной жизни. Культурологическая концепция перевода связана с проблемами межкультурной коммуникации, когда при общении между коммуникантами участвует посредник-переводчик. Перевод рассматривается как частный случай межкультурного общения. С этой точки зрения выделяются следующие виды межкультурной коммуникации: двуязычная межкультурная коммуникация, когда общение протекает в кодах разных знаковых систем с участием переводчика; одноязычная межкультурная коммуникация; межсубкультурная коммуникация, которая осуществляется представителями субкультур одной культуры (о межкультурной коммуникации см. также [Садохин 2004]).

Модели машинного (автоматического) перевода

Трансформационная теория и модель перевода. В данном случае, как видим, соединены теория и модель. Трансформационная теория перевода основывается на изучении процесса перевода как преобразования единиц структуры исходного языка в единицы перевода языка переводящего. Ее разрабатывали И. И. Ревзин и В. Ю. Розенцвейг в нашей стране и Ю. Найда за рубежом. В основе этой теории лежат идеи порождающей (трансформационной) грамматики американского ученого Н. Хомского [Ревзин и др. 1964]: все существующие грамматически правильные высказывания порождаются ограниченным числом ядерных (базисных) структур предложений. Любой передаваемый на исходном языке текст может быть транс-

формирован в ядерные (базисные, глубинные) структуры. Для передачи этого текста на другом языке его необходимо снова трансформировать, т.е. развернуть в речь из глубинных ядерных структур в поверхностные (предложения) по законам и правилам переводящего языка.

Заметим, что эта теория предполагается единой и для «ручного» перевода, и для перевода машинного. Н. Хомский является математиком, и его подход типичен для математического, абстрактного взгляда на язык как на средство кодирования мысли. Ядерные структуры предполагаются едиными для любых человеческих языков, развертывание смысла (содержания) высказывания — происходящим по законам логики. Разницу между смыслом и содержанием мы уже иллюстрировали. Недостаток трансформационной теории перевода Л. Л. Нелюбин справедливо видит в том, что она не имеет экспериментальной основы, т.е. оторвана от объекта науки о переводе. Эта теория исходит из признания субъективного смыслового кода, описанного в трудах психологов, прежде всего в работах Н. И. Жинкина, хотя и сводит этот код к лингвистическим ядерным структурам. В трансформационной теории перевода наметился отход с чисто лингвистических позиций в сторону литературоведческих, замечает Л. Л. Нелюбин, что превращает переводчика в литератора, и признает за ним творческое начало. Наверное, можно и так оценивать эту теорию, которая с точки зрения автоматизации перевода и создания перевода машинного не принесла существенных результатов, хотя, как сказано выше, попыток ее осуществления в этом направлении было достаточно.

Теория машинного перевода. Теперь рассмотрим некоторые проблемы теории и моделирования машинного перевода. Выше, в соответствующем разделе, мы уже кратко излагали основные этапы становления и развития машинного (автоматического) перевода. В этом параграфе остановимся на основных теоретических положениях, с помощью которых так или иначе моделировался процесс перевода. Изложим точку зрения профессора Л. Л. Нелюбина на эту проблему.

В связи с возникновением в середине XX в. специфического объекта переводоведения — перевода с использованием компьютера, так называемого машинного (автоматического) перевода — появляется потребность в особой теории, изучающей проблематику, а именно теории машинного перевода, которая нашла свое отражение и воплощение, с одной стороны, в многочисленных описательных трудах «создателей» так называемых «бумажных» алгоритмов машинного перевода, а с другой существенной стороны — в прагматических промышленных системах, описанных и предъявленных в трудах Всесоюзного центра переводов научно-технической литературы и документации при Академии наук под руководством и непосредственным участием Ю. Н. Марчука, в работах Всесоюзной группы «Статистика

речи» под руководством и авторским участием профессора Р. Г. Пиотровского, группы Ю. А. Моторина (основоположников машинного перевода в СССР), в монографических работах и вероятностных экспериментальных и промышленных системах К. Б. Бектаева, Л. Н. Беляевой, И. К. Бельской, А. В. Зубова, Г. Э. Мирана, Л. Л. Нелюбина, В. И. Перебейнос, В. А. Чижиковского и др.

Фактически теория машинного перевода начала зарождаться после проведенного первого в мире опыта машинного перевода с русского языка на английский, который был осуществлен в январе 1954 г. в Нью-Йорке, в Джорджтаунском университете, под руководством Леона Достерта (так называемый Джорджтаунский эксперимент). В декабре 1955 г. в Москве, в Академии наук, был проведен первый в мире опыт машинного перевода с английского языка на русский под руководством И. С. Мухина, Л. Н. Королева, С. Н. Разумовского и Д. Ю. Панова. Лингвистический алгоритм был разработан И. К. Бельской.

В 1956 г. в Москве была организована Лаборатория машинного перевода под руководством Ю. А. Моторина, в состав группы входил молодой ученый Ю. Н. Марчук, впоследствии возглавивший Всесоюзный Центр переводов. В 1958 г. система машинного перевода группы Ю. А. Моторина осуществила перевод с английского языка на русский текстов общественно-политической (газетной) и научно-технической тематики любого объема. Это был первый в мире промышленный, экономически выгодный машинный перевод [Котов и др. 1983].

Далее читаем у Л. Л. Нелюбина. «Если говорить об упомянутых выше «бумажных» алгоритмах, то следует привести высказанное по этому поводу суждение Ю. Н. Марчука. Ю. Н. Марчук справедливо считает, что машинному переводу в СССР немного не повезло, потому что идея машинного перевода сразу же привлекла внимание некоторой группы теоретиков, которые усмотрели в ней возможность захватить главенствующее положение в языкознании, не прилагая при этом никаких особых усилий к реальному доказательству слишком абстрактных теорий.

Дело в том, что машинный перевод, пишет Ю. Н. Марчук, как и всякий перевод, чрезвычайно интересен тем, что здесь есть возможность сразу же проверить истинность и объяснительную силу описывающих перевод моделей, ибо составленный алгоритм и проведенный перевод сразу же покажут эффективность предложенных творческих рассуждений.

Однако «теоретики» пошли по другому пути, выдвинув лозунг «машинный перевод без перевода, без машин, без алгоритмов». Таким образом, защищалась идея осуществления перевода только за счет декларативных знаний, когда лингвист дает лишь описание фактов языка, а построение алгоритма перевода передается в руки математика и программиста. Такая спекуляция на непонимании линг-

вистических концепций в языке дала возможность упомянутой группе «теоретиков» занять ведущее положение и существенно препятствовать развитию прикладных и практических работ в области МП.

Эта спекуляция в значительной степени также велась на уровне вычислительной техники. «Теоретики» убедительно говорили, что современные (того времени) компьютеры не в состоянии реализовать такие универсальные идеи, как «текст-смысл-текст», а более совершенные машины могут все сделать. Но сейчас мы имеем гораздо более совершенные компьютеры, чем мы могли даже представить, а до реализации фантастических идей «бумажных алгоритмов» так же далеко, как и раньше.

Вред от вышеназванных концепций также и в том, справедливо считает Ю. Н. Марчук, что реальные усилия исследователей отвлеклись от нужных направлений в теории и практике машинного перевода, что явилось следствием отставания лингвистического обеспечения современных систем перевода в рамках конкретных языковых пар. Поэтому сейчас важно осознать и преодолеть это отставание как в теории прикладной (компьютерной) лингвистики, так и в практике сопоставительного описания переводных соответствий в конкретных языковых парах.

Существующий в настоящее время позитивный подход в теории и практике машинного перевода позволяет создавать достаточно эффективные системы практически пригодного машинного перевода и компенсировать спровоцированное отставание» [Нелюбин 2007].

Глава 9

МОДЕЛИРОВАНИЕ В КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

В данной главе мы рассмотрим проблемы, связанные с изучением и использованием в целях компьютерной (прикладной, вычислительной) лингвистики языковых единиц, больших, чем морфема, слово, словосочетание. Речь пойдет о предложении, высказывании, синтагме и пр., т.е. о лингвистических единицах, выражающих то, что можно определить как некоторую законченную мысль. Если все предыдущие единицы относились в общем случае как бы к имени, т.е. являлись результатом номинации, простого (или сложного) наименования объектов высказывания и мысли, то теперь мы переходим к единицам, содержащим предикативность, т.е. есть имя, о котором что-то сказано, констатировано наличие признака, действия, состояния и пр. В языке подобная мысль выражается высказыванием, более или менее законченным. Формализация смысла и содержания высказывания тесно связана с его пониманием как первым

этапом перевода. Уровни синтаксического и семантического анализа в прикладных целях определяются в первую очередь мышлением человека. Поэтому в данном разделе речь пойдет о соотношении языка и мысли, об интеллекте, о моделировании интеллекта и создании систем «искусственного интеллекта». Мы не будем затрагивать философских аспектов этой проблемы. Нас интересуют прежде всего прикладное, лингвистическое содержание дела, возможности использования компьютера для решения задач, так или иначе связанных с мышлением.

Моделирование языковых сущностей и человеческого мышления

Связь языка с мышлением

Связь языка с мышлением — сложный философский вопрос, над которым работают ученые с древних времен до наших дней. Мы исходим из предположения, что такая связь есть, она является решающей и принимает такие формы, которые можно использовать для более или менее успешного формального описания высказывания, предложений и текста. «В том, что язык является средством формирования мысли, нет никакого сомнения. После В. Гумбольдта эта идея для всех стала привычной. Однако при конкретном исследовании языка и мысли большинство авторов обнаруживают между ними такое глубокое различие, что невольно начинают рассматривать язык и мысль в виде двух параллельных взаимосвязанных потоков. И какие бы при этом ни произносились слова о неразрывном существовании языка и мысли, представление об оголенном существующей хотя и в какой-то связи с языком, но лишь рядом с ним и вне его» [Кривоносов 2001].

А. Т. Кривоносов, рассматривая проблему частей речи с теоретической точки зрения, включает ее в структуру «действительность-мышление-сознание-язык». Понятие «мышление» в языкознании, логике, психологии, философии настолько неопределенно и так запутанно, говорит он, что под «мышлением» некоторые исследователи иногда понимают чуть ли не противоположное. Прежде всего потому, что само содержание термина «мышление» расплывчато и не имеет четких границ [Кривоносов 2001].

По степени, глубине, силе абстракции и обобщения необходимо различать, по А. Т. Кривоносову, два уровня мышления, или две ступени познания: чувственное (непосредственное, наглядное, эмпирическое, техническое мышление, осуществляющееся на первой ступени абстракции) в формах ощущений, восприятий, представлений, которые не требуют языковой опоры (первая сигнальная система, по

И. П. Павлову), и абстрактное, логическое, теоретическое мышление, осуществляющееся на второй ступени абстракции (соответственно, по И. П. Павлову, вторая сигнальная система). Это мышление реализуется в логических формах понятий, суждений, умозаключений, обязательно выраженных в формах естественного языка — в словах (понятиях) и предложениях (суждениях, умозаключениях). Логическое мышление также может быть выражено и вне форм языка. Это происходит тогда, когда мы думаем, рассуждаем про себя, вспоминаем о чем-нибудь, не прибегая к формам языка. Именно безъязыковое (авербальное) мышление и является, в сущности, типичной формой человеческого мышления.

Две ступени познания — чувственное созерцание, осуществляющееся посредством органов чувств, и абстрактное, логическое мышление, обобщающее данные чувственного познания, — образуют единый процесс мышления, в результате чего происходит постоянное превращение элементов, или континуума, чувственного мышления в абстрактное и абстрактного в чувственное. Далее различаются два уровня движения логического мышления: «процесс мышления» (динамика мышления) и «сознание» (статика мышления). В чувственном и логическом мышлении мы обнаруживаем сложное взаимодействие между реальной действительностью и мозгом, а в логическом мышлении — также и с языковыми знаками.

А. Т. Кривоносов далее в этой работе критикует «когнитивную лингвистику» за то, что она только запутывает старые проблемы языкознания. Для того чтобы разобраться в основных понятиях мышления, рассмотрим общепринятые точки зрения на такие предметы, как искусственный интеллект и процесс мышления у человека.

Элементы системы искусственного интеллекта

В процессе умственной целенаправленной деятельности человек формирует подлежащие решению задачи, отыскивает правила, учитывает разного рода обстоятельства, наконец, вырабатывает решение, отбрасывая ненужную информацию. Все эти элементы мыслительной деятельности должны как-то закладываться в программу компьютера, если мы хотим, чтобы он обладал «мышлением». Модель функционирования системы искусственного интеллекта может быть представлена следующим образом (рис. 11).

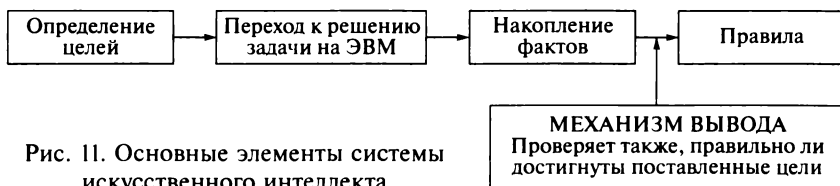


Рис. 11. Основные элементы системы искусственного интеллекта

Приведенная на этом рисунке структура компьютерной модели искусственного интеллекта основана на раздельном функционировании различных компонентов системы [Ефимов, Фролов 1991]. Этот самый важный фактор и делает программу искусственного интеллекта гораздо более эффективной по сравнению с другими, обычными компьютерными программами. Эта концепция структурной автономности очень важна и с другой точки зрения. Поскольку элементы программы отождествляются с «составными частями» человеческого мозга, можно заложить в нее сам способ мышления высококвалифицированного специалиста по любой отрасли знания. Если можно определить, что «делается» в его мозге на каждой стадии этого процесса, мы можем легко подыскать эквивалентный участок программы, который соответствует такому же участку в человеческом мышлении. Поскольку достижение целей является задачей любой системы искусственного интеллекта, первым шагом в ее создании должно быть точное определение этих целей. Необходимо знать, какого рода проблему вы хотите решить, и суметь ее описать, пользуясь конкретными обозначениями и терминами, прежде чем дать соответствующее задание компьютеру.

Как мыслит человек

Под системой с названием «искусственный интеллект» в практическом плане обычно понимают компьютерную программу, способную «думать» и решать так называемые «творческие» задачи.

Для реализации «искусственного интеллекта» нужно прежде всего изучить, как мыслят люди, когда им надо принять какое-либо решение или решить ответственную проблему. Следует выделить в мыслительном процессе основные стадии, что и позволит в дальнейшем разработать компьютерную программу, способную решать задачи с использованием тех же стадий мыслительного процесса. Искусственный интеллект, таким образом, обеспечивает, по крайней мере, простой структурный подход к разработке весьма сложных программ, позволяющих решать творческие задачи весьма разного уровня.

Стандартная компьютерная программа способна обеспечить ответы только на те вопросы, для которых она специально создана. Если требуется изменить стандартную программу для пополнения ее новой информацией, то приходится ее тщательно просматривать, чтобы найти место для новой информации. Это не только отнимает время, но и может нанести вред программе, которая может теперь давать ошибки из-за влияния на другие ее части. Искусственный интеллект, как и предполагает название, действительно делает компьютер способным думать. Он имитирует основной человеческий процесс обучения, при котором происходит прием информации и ее усвоение для дальнейшего использования. Человеческий мозг способен воспринимать все новые и новые знания без изменения про-

цессов жизнедеятельности и без функциональных нарушений различных отделов головного мозга. Программа искусственного интеллекта действует почти таким же образом.

Разумеется, стандартная программа может выполнить почти все, что и программа искусственного интеллекта, но составить ее нельзя так же легко и быстро, говорят Н. Н. Ефимов и В. С. Фролов. В обоих типах программ есть блоки, независимые в смысле выполнения разработанных пользователем функций. Но программа искусственного интеллекта обладает одной отличительной чертой, которая приравнивается к черте человеческого интеллекта. Любое звено в ней может быть изменено и модифицировано без вредного воздействия на структуру всей программы. Такая гибкость обеспечивает большую эффективность программирования и способность понимания пользователем всех шагов ЭВМ. Иными словами, компьютер обретает «интеллект».

Искусственный интеллект

Литература по искусственному интеллекту в настоящее время достаточно обширна [Поспелов 1983, Шабанов-Кушнаренок 1984, Потапова 2002, Зубова 2001, Рябцева 2002 и др.]. В данном параграфе мы рассмотрим некоторые вопросы моделирования мышления в целом, прежде чем в дальнейшем перейти к моделированию лингвистических способностей человека в части синтаксиса и семантики.

Модель механизма мышления

Любого человека легко поставить в ситуацию, когда он не сможет адекватным образом отреагировать на происходящее (выполнить нужное действие, правильно ответить на вопрос и пр.). Умение правильно и своевременно реагировать в самых различных ситуациях у человека и животных приобретает лишь благодаря способности к обучению, без которой было бы невозможно развитие человека [Вайнцвайг и др. 1987]. Эта способность тесно связана с требованием работы в реальном времени, поскольку в результате обучения появляется возможность не только решать новые, ранее недоступные задачи, но и быстро решать те, которые раньше решались медленно.

Оптимальная с точки зрения выживаемости организма модель механизма мышления должна удовлетворять следующим условиям:

- работать в реальном времени;
- обладать способностью к обучению;
- максимально полно использовать поступающую извне информацию как на этапе принятия решений, так и на этапе обучения;
- иметь полную память о прошлых событиях и обладать способностью непрерывного обобщения поступающей информации.

На рис. 12 изображена такая модель механизма мышления.



Рис. 12. Модель механизма мышления

Блоки всей этой системы выполняют следующие функции.

Информацию о внешнем мире и о своих внутренних состояниях система получает с помощью рецепторов, а посредством эффекторов она выполняет те или иные внешние действия. Ее поведение в общем случае управляется двумя функциональными блоками: сенсорным анализатором и ассоциативной памятью-процессором. Сенсорный анализатор — устройство, производящее обработку входной (сенсорной) информации, основанную на жестких, необучаемых механизмах, — осуществляет согласование информации, получаемой от различных органов чувств, и обеспечивает константное восприятие важных свойств и пространственно-временных характеристик предметов. В результате своей работы он реализует следующие функции:

- формирует первичные оценки качества ситуаций — первичные эмоции (меру удовольствия или неудовольствия и пр.);
- при выполнении соответствующих условий формирует команды безусловно-рефлекторных (инстинктивных) действий;
- при выполнении определенных условий формирует те или иные цели (построить гнездо, добыть пищу и пр.);
- строит описание текущего состояния организма в терминах его ощущений, производимых действий, поставленных целей и испытываемых эмоций, которые, попадая в поле внимания, всегда запоминаются, причем все те и только те, которые проходят через поле внимания.

В процессе ассоциации осуществляются: формирование понятий и законов, постановка задач, решение задач и формирование вторичных эмоций. Функциями ассоциативной памяти-процессора в основном и определяется весь мыслительный процесс.

Слово «осознать» можно рассматривать как синоним выражения «обратить внимание на ассоциируемый объект», а слово «сознание» — как синоним «память об ассоциации». В частности, если вспоминается (т. е. вновь поступает в поле внимания и заносится в память с модальностью «из памяти») только что виденный объект, говорят

об осознании факта видения этого объекта; если образ конкретной собаки вызывает в памяти ассоциацию с ранее сформированным обобщенным понятием «собака», говорят об осознании того, что в данный момент видится собака.

Ассоциативное построение понятий

В памяти имеется информация двух типов: основные данные, к которым относятся описания ощущений, эмоций, выполняемых действий, законов, задач и путей их решения, — словом, все, что проходило когда-то через поле внимания, сохраняемое в неизменном виде, а потому всегда доступное сознанию; соответствующие основным данным вспомогательные параметры, такие, как величины ассоциации, важности, надежности предсказания и пр., постоянно перевычисляемые в процессе ассоциаций и используемые в процессе обучения, постановки и решения задач, возникновения вторичных эмоций, переключения внимания и пр.

Ассоциация изначально ориентирована на формирование и работу с выражениями некоторого языка, на котором описываются понятия (предикаты от ситуации), а также законы и задачи. Такой язык должен удовлетворять следующим требованиям. Ассоциация должна быть в равной мере применима как к исходным описаниям, так и к сформированным вариациям этого языка. В частности, должна реализовываться возможность выяснения применимости к конкретному описанию тех или иных обобщений. Кроме того, выражения языка должны достаточно просто формироваться в процессе ассоциации.

От выбора конкретного языка и критериев формирования его выражений в значительной мере зависит эффективность обучения и работы системы. Предварительно в рассматриваемой модели в качестве такого языка рассматривается расширение языка регулярных выражений.

Понятия и законы внутреннего языка в системе предполагается формировать рекурсивно, когда сложные понятия строятся на основе ранее построенных простых понятий. Идея состоит в том, что понятия ограниченной сложности могут строиться в процессе последовательных ассоциаций непосредственно в тех же местах памяти, где хранятся исходные описания.

В результате ассоциации в каждом из ассоциаторов, принадлежащих второму этажу памяти, пропорционально энтропии распределения его весов, величине ассоциации и надежности предсказания последующих элементов памяти происходит относительное увеличение весов тех способов интерпретации, которые принимали участие в ассоциации, т. е. принадлежат ее следу.

Последовательность значений ассоциаторов на некотором отрезке содержимого памяти представляет собой как бы текущее значение обобщенного описания ситуации — своеобразное нечеткое понятие,

эквивалентное некоторому множеству предикатов от ситуации с заданным на нем распределением весов. В процессе последующих ассоциаций на отдельных участках памяти будет происходить постепенное контрастирование распределения весов различных способов интерпретации, приводящее к выделению среди них наиболее предпочтительных.

В сформированном виде, когда состояния ассоциаторов определяют лишь единственную возможную последовательность типов интерпретации, обобщенное описание ситуации соответствует тому, что обычно называют понятием, и формально эквивалентно одиночному предикату от ситуации.

Одновременно с построением понятий происходит формирование законов, которые строятся в виде утверждений, имеющих вид причинно-следственной связи.

Основные принципы работы системы и организации ее поведения

Существуют три уровня организации поведения:

– безусловно-рефлекторный (инстинктивный), характеризуемый тем, что отсутствует использование каких-либо сведений из основной памяти;

– условно-рефлекторный (бессознательный), характеризуемый тем, что при организации поведения хотя и используются сведения из основной памяти, но никакие промежуточные результаты процесса этой организации не осознаются, т.е. не попадают в поле зрения и не запоминаются;

– сознательный уровень, основанный, как правило, на многократном (циклическом) использовании памяти, когда результаты работы каждого шага цикла, попадая в поле внимания и вызывая в памяти дальнейшие ассоциации, запоминаются.

Подробнее о сознательном поведении авторы говорят следующее. Подобно тому как в обычной вычислительной машине не любая функция может быть реализована без использования дополнительных полей памяти, не любая поведенческая задача может быть решена на условно-рефлекторном уровне. Существуют сложные задачи, решение которых требует многократного (циклического) обращения к памяти через поле внимания, что соответствует сознательному уровню организации поведения. Это происходит, например, в случаях, когда задачу приходится разбивать на подзадачи. Пример: сложение чисел, когда каждый следующий разряд суммы последовательно проходит через сознание, требуя для своего вычисления новой ассоциации с таблицей сложения. На сознательном уровне решаются задачи, для которых в памяти не хранятся готовые решения, и их приходится искать. Узловые моменты этого поиска, проходя через сознание и вызывая дальнейшие ассоциации, запоминаются, во-

первых, для того, чтобы не приходилось повторять уже сделанных ходов, а во-вторых, чтобы найденным решением можно было потом воспользоваться на условно-рефлекторном уровне.

В основе сознательного поведения лежит процесс возникновения желаний.

Желание — это осознанное побуждение, т. е. фиксируемое в памяти так, что его можно потом воспринимать, стремление сохранить или избавиться, достичь или избежать какого-либо свойства текущей ситуации. Оно определяет некоторую долговременную (конечную или промежуточную) цель поведения. В отличие от побуждений, величина которых постоянно зависит от текущей ситуации и с ее изменением может резко изменяться, величина желания, будучи занесенной в память в виде данных основного типа, служит постоянным стимулом в поведении, т. е. влияет на процесс формирования побуждений до тех пор, пока в результате осознания некоторого побуждения, например, не появится новое желание, противоречащее первому.

Эта общая модель человеческого поведения, на основе которой объясняется феномен мышления и выделяются основные компоненты модели, может служить некоторой основой для рассмотрения того, как модель интеллекта использует знания и умения, в том числе и лингвистические. Достаточно сказать, что все потоки информации внутри модели между ее составными частями могут быть представлены в символах некоторого языка, необязательно естественного, но тем не менее образующего знаковую систему, совершенно подобную человеческому языку. Степень сложности передаваемой информации определяет степень сложности используемых языковых средств. Не касаясь здесь теории языкового знака, можно безошибочно утверждать, что символичный язык внутри модели имеет как план выражения, так и план содержания и что он может изучаться как некоторая самостоятельная сущность.

Представление знаний

Следующим важным этапом рассуждения будет определение того, что именно передается при обмене информацией между компонентами мыслительной структуры. Каковы способы представления «содержания» человеческого мышления, способы представления того, что мы называем «знаниями»?

В общем виде под «знаниями», используемыми в системах моделирования интеллекта, понимаются особым образом организованные данные. Заметим, что здесь возникает некоторый замкнутый круг («знания» = «данные» или «сведения»), который, однако, не мешает интуитивному представлению о том, что такое знания. Для того чтобы знания могли использоваться в системе, они должны быть формализованы, для чего применяется специальный матема-

тический аппарат. Разные способы представления знаний отличаются друг от друга видом и характером применяемого математического аппарата.

В большинстве случаев в основе представления знаний лежит логическая форма выражения. Элементарной единицей чаще всего выступает суждение, или утверждение [Кривоносов 1996]. Знание некоторой системы или определенной предметной области, складывается из совокупности утверждений относительно этой системы или области. Для представления знаний используется аппарат формальной логики, в частности исчисление предикатов, позволяющий не только фиксировать множество заданных суждений, но и выводить из них с помощью специальных правил некоторые новые суждения.

Обычно выделяют два способа представления знаний: декларативный и процедурный. **Декларативное представление знаний** есть множество утверждений, в значительной степени независимых от того, где их можно использовать. Оно состоит из аксиом, всех выведенных к тому времени теорем (в широком смысле слова) и множества операторов, представленных правилами вывода. Описывая некоторый факт или явление декларативно, мы можем обрисовать его структуру в терминах определенных признаков, указать на связь с другими аспектами или явлениями, назвать связанные с ним прагматические аспекты. Однако в этих знаниях не будет указаний на те реальные действия, которые должны породить эти знания. В этом смысле декларативные описания пассивны. Примером декларативных описаний может служить научная классификация явлений, например систематика растений, основы которой изложены К. Линнеем. Там установлены системы признаков, определяющих различные виды и подвиды растений, указаны родо-видовые и классов-элементные отношения, однако какие-либо действия не обозначены. Есть, правда, классификации, в которых факты или явления определены по признаку принадлежности к какой-либо ситуации. В этом случае уже подразумеваются некоторые действия. Д. А. Поспелов отмечает, что в этом случае уже в декларативных описаниях содержится некоторая информация о процедурах [Поспелов 1983].

При **процедурном представлении информации** знания даются в виде процедур их преобразования для данной предметной области. Пример процедурного представления знаний — инструкция, которую можно прочитать в кабине междугороднего телефона-автомата. Сведения о признаках сведены тут к минимуму и присутствуют только тогда, когда наличие или отсутствие определенных свойств приводит к остановке процедуры (например, признак «юбилейности» или «гнутости» монет). Процедурные описания, подобно декларативным, могут вступать между собой в определенные отношения. Так, сценарии или семантические сети специального типа могут связывать процедуры друг с другом. Связи эти могут быть иерархическими, ситуативными или иного типа.

Противопоставление декларативного и процедурного способов представления знаний не означает их коренной несовместимости; на самом деле, как видно из уже сказанного выше, в декларативном представлении есть элемент процедуры, а в процедурном — составляющие декларативного описания, например признаки. Еще большее слияние этих двух способов имеет место в таких, например, представлениях знаний, как семантические сети или фреймы.

Семантические сети и фреймы

Семантическая сеть представляет собою граф, в вершинах которого фиксируются наименования данных, а дугам соответствуют отношения, устанавливаемые между ними. Если некоторая предметная область может быть представлена набором двухместных предикатов, группирующихся в блоки, представляющие собой множество утверждений о некотором объекте, то она может быть представлена и в виде семантической сети. Бинарные предикаты допускают представление с помощью графов, где дугам соответствуют предикаты, а вершинам — аргументы. Совокупность таких структур, представленных в виде графа, и образует **сеть**, называемую семантической. В реальных системах применяются сети различной степени сложности, в некоторых из них дугам могут соответствовать формулы или сложные математические модели.

Сочетание приемов из декларативного и процедурного представлений с принципами, используемыми в семантических сетях, привело к созданию другого типа семантического представления — **фреймов**. Их особенностью является то, что здесь используются модульные структуры, представляющие собой некоторые относительно самостоятельные блоки или единицы.

Фрейм можно рассматривать как один из видов воспринимаемого объекта, вид в перспективе, который (вид) может быть представлен как некоторый граф [Шемакин 1985]. Вершина такого графа соответствует наименованию объекта, а подчиненные вершины — элементам этого объекта, которые наблюдатель видит с определенной точки зрения. Изменение положения объекта относительно наблюдателя приводит к формированию других фреймов, поскольку видимыми становятся уже другие элементы объекта. Невидимые элементы не исчезают из памяти, а запоминаются, что находит выражение и в формальной записи новых фреймов. Группа связанных между собою фреймов образует систему.

Таким образом, фрейм — это некоторая структура, содержащая сведения об определенном объекте и выступающая как целостная и относительно автономная единица знания. В долговременной памяти человека хранится большой набор систем фреймов, которые используются, например, при распознавании зрительных образов. Представление о фреймах получило дальнейшее развитие и интерпрета-

цию. Фрейм отождествляется с некоторой стандартной, однотипной ситуацией, включающей в себя конкретное множество отдельных однородных ситуаций. В зависимости от классов ситуаций различают фреймы визуальных образов, фреймы-сценарии, семантические фреймы и т.п. Во фреймах, используемых для систем ИИ, с целью представления знаний можно выделить несколько уровней вершин или узлов, иерархически связанных между собой. Таким образом, структура фрейма включает в себя три основных типа данных: понятие (название фрейма), характеристика (название терминала), значение характеристики (заполнители терминала). Можно считать, что во фрейме реализуются некоторые общие принципы, свойственные организации баз данных, где в качестве единиц выделяются объекты, характеристики и их значения, а также семантическим сетям, в которых различаются абстрактный и конкретный уровни. В то же время фрейм имеет и свою специфику.

Каждую конкретную базу данных или семантическую сеть можно рассматривать как некоторую макроединицу, состоящую из микроединиц, в качестве которых выступают наименования данных или вершин сети. Фрейм является самостоятельной единицей, имеющей черты как макро-, так и микроединиц. Статус его как макроединицы определяется тем, что он также является сетью. Но эта сеть локализована данными, необходимыми для определенных объектов или ситуаций. Будучи сетью, фрейм включает в себя и микроединицы в виде наименования вершин. Но и сам он выступает в роли микроединицы, когда входит в состав некоторой системы или сети фреймов. Следовательно, фрейм представляет собой единицу, промежуточную между микро- и макроединицами. Отличительной его особенностью является также и то, что не все его элементы заданы априорно, как это имеет место, например, в семантических сетях. Он допускает различные варианты заполнения терминалов, что придает этому способу представления знаний необходимую гибкость и адаптивность.

Рассматривая разные способы представления знаний о мире, нужно отметить, что в целом ни один из них не является сам по себе достаточным, универсальным и позволяющим игнорировать все другие. Выбор того или иного способа зависит от конкретной задачи и предметной области.

Знание как объект моделирования

Лингвистический аспект представления знаний

Совершенствование систем общения прежде всего связано с решением вопроса о том, что такое знание, выступающее в них в качестве объекта моделирования. Знание может рассматриваться в пси-

хологическом, гносеологическом, лингвистическом и других аспектах. Рассмотрим лингвистический аспект представления знаний.

А. И. Новиков полагает, что лингвистическое представление знаний лучше всего проявляется при рассмотрении исходного понятия «текст». В нем находит отражение связь между конкретными элементами знания и конкретными языковыми средствами, используемыми для их выражения. Подход к представлению знаний, базирующийся на анализе текстов, относящихся к определенной предметной области, включает в себя два основных этапа: выделение и эксплицитное представление структуры содержания отдельного текста; объединение отдельных структур содержания конкретных текстов и построение на этой основе общей системы знания соответствующей предметной области.

Текст рассматривается не только как готовый продукт, но и как процесс. Это объясняется тем, что содержание текста формируется в интеллекте человека как результат осмысления составляющих его языковых единиц и понимания текста в целом. Следовательно, оно может быть выявлено при учете основных закономерностей восприятия и понимания текста, осуществляющегося на разных уровнях его организации и проходящего ряд этапов. Для определения структуры содержания необходимо выявить такие единицы мыслительного плана, которыми мышление оперирует на заключительных этапах понимания. Основная сложность заключается в том, что понимание непосредственно не заканчивается на анализе текста. Мыслительное образование, возникшее в результате осмысления языковых средств, может подвергаться дальнейшему осмыслению уже независимо от данного конкретного текста. Поэтому собственно содержанию текста должно соответствовать такое мыслительное образование, которое, с одной стороны, образуется на наиболее глубинном этапе понимания, а с другой, — является непосредственным результатом воздействия на интеллект совокупности языковых средств, составляющих данный текст [Котов и др. 1987].

Такое мыслительное образование можно назвать **внутренней формой текста**, которая является формой существования его содержания и находится в определенном отношении с внешней формой текста.

Под **внешней формой** понимается совокупность языковых средств, включающая их содержательную сторону и реализующая замысел автора. Это то, что дано для непосредственного восприятия и что должно быть осмыслено и понято. То, что понимается, составляет внутреннюю форму. Это такое мыслительное образование, которое возникает в интеллекте человека и соотносится с внешней формой не поэлементно, а в целом соответствует всей совокупности данных языковых средств.

Переход от внешней формы текста к его внутренней форме и составляет сущность процесса понимания.

Понимание текста

Понимание сопровождается интенсивным преобразованием и перестройкой исходной (внешней) формы текста, осуществляющейся в результате мыслительной аналитико-синтетической деятельности и приводящей к свертыванию текста. Результатом преобразования и свертывания текста является совокупность «контекстных объединителей» (смысловые знаки, опорные пункты), каждый из которых замещает в мышлении определенный фрагмент текста. Преобразование исходной формы текста проходит ряд этапов и заключается в переводе его на «язык» внутренней речи, которая представляет собой центральное звено, где совершается качественный скачок от «внешних кодов» языка к «внутренним кодам» интеллекта. Перевод на «язык» внутренней речи не является конечным этапом перехода к содержанию. Этот переход продолжается и на следующих этапах. Содержание текста соответствует денотативному уровню отражения, т. е. знанию о том фрагменте действительности, о котором сообщается в тексте.

Содержание текста представляет собой динамическую модель ситуации, заданную всей совокупностью средств, составляющих внешнюю форму текста. Основной единицей содержания является денотат, понимаемый как мыслительное образование, соответствующее свернутым образам моделей предметов или предметных комплексов, о которых сообщается в тексте. Содержание текста включает в себя не только те денотаты, которые имеют в тексте непосредственное выражение, но и подразумеваемые, имплицитные денотаты.

Динамическая модель ситуации, задаваемая текстом, представляет собой систему денотатов, способ организации которой не базируется на грамматических и логико-композиционных закономерностях текста, а подчиняется логике предметных отношений, связывающих денотаты между собой. Между элементами внешней формы и единицами содержания нет однозначного соответствия и непосредственного перехода, поскольку содержание соответствует в целом всей совокупности элементов внешней формы. Между внешней и внутренней формой существует опосредованная связь, устанавливаемая мышлением в процессе восприятия и понимания текста.

Содержание текста не существует вне процесса его восприятия и понимания, поскольку оно каждый раз формируется в интеллекте партнера по коммуникации как результат воздействия на него соответствующих сигналов, в качестве которых выступают языковые средства. Оно включает в себя новую информацию, которая находится в определенном отношении с информацией репродуктивной.

Денотативный анализ текста

Денотативный анализ текста завершается эксплицитным представлением его содержания в виде графа денотатной структуры, вер-

шинам которого соответствуют имена денотатов, а ребрам — отношения между ними. По А. И. Новикову, основные этапы, через которые проходит процесс построения денотатной структуры, реализованный в соответствующих компьютерных программах в виде действующей системы смыслового анализа текста, следующие:

- определение имен денотатов в тексте;
- выделение наименований «ключевых» денотатов;
- установление внутренних связей каждого ключевого денотата с другими денотатами данного текста;
- определение предметных отношений между денотатами;
- перестройка структуры связей между денотатами, задаваемых текстом, в соответствии с предметными отношениями этих денотатов;
- выявление имплицитных денотатов и их предметных отношений;
- формирование целостной структуры содержания с учетом предметных отношений денотатов и их места в этой структуре в соответствии с тем, в какой роли (подтемы, субподтемы и пр.) они выступают в тексте.

Определение структуры содержания текста является предпосылкой к решению проблемы представления знания, но не решает ее полностью. Объясняется это тем, что такие структуры представляют собой отдельные фрагменты знания, содержащиеся в конкретном тексте, в то время как требуется построение системы знания, относящейся к определенной предметно-тематической области в целом. Это может быть достигнуто за счет интеграции отдельных денотативных структур в некоторую целостную структуру.

Данная концепция моделирования знаний с целью извлечения смысла из текста получила практическую реализацию и используется для решения конкретных задач содержательного анализа и связей текстов [Новиков 2002].

Лингвистический алгоритм, нечеткие множества, порождение речи

Особенности лингвистического алгоритма

Моделирование в компьютерной лингвистике, как правило, подразумевает построение алгоритма обработки текста на естественном языке с теми или иными целями. Что представляет собой алгоритм, используемый в модели для решения конкретной задачи обработки текста?

Известно, что понятие «алгоритм» восходит к имени арабского ученого Аль-Хорезми, который предложил формальную процедуру, описывающую четыре арифметических действия в десятичной си-

стеме. Под алгоритмом понимается конечный набор правил, позволяющих чисто механически решать любую конкретную задачу из некоторого класса однотипных задач. Интуитивно-содержательное определение алгоритма содержит три его свойства [Советская энциклопедия 1980]:

- детерминированность: метод вычисления должен быть настолько точен, чтобы не оставалось места произволу. Этот метод можно сообщить в виде конечного числа указаний любому лицу;

- массовость: алгоритм служит для решения не какой-либо одной конкретной задачи, а целого класса задач;

- результативность (направленность) алгоритма: алгоритмическая процедура, примененная к задачам определенного типа, через конечное число шагов должна останавливаться и давать возможность получить определенный результат.

Алгоритмы бывают численные и логические. Пример численного алгоритма — арифметические действия над числами; логический алгоритм — поиск пути в лабиринте. В первом случае алгоритм работает с числами, во втором — с высказываниями. При работе с высказываниями требуется, во-первых, чтобы были четко определены границы, отделяющие одно высказывание от другого, и, во-вторых, высказываниям должны быть сопоставлены однозначные характеристики, например: истинно данное высказывание или ложно (1 или 0).

Когда мы пытаемся применить алгоритм к лингвистическим задачам, мы видим, что объекты, к которым он применим, имеют особую качественную природу. Возьмем, например, множество слов естественного языка. Конечно ли оно? Можно ли ему сопоставить какие-то количественные характеристики? Ответ на этот вопрос дают разные объемы разных словарей одного и того же языка. Всегда можно пополнить исходный объем слов данного множества хотя бы одним словом. Если мы будем вводить ограничения на систему признаков, по которым причисляем тот или иной лингвистический объект к данному множеству, то все равно будет очень трудно получить действительно конечное множество, например все существительные или все глаголы данного языка.

Если мы рассмотрим проблему плана содержания, то и здесь увидим определенную нечеткость в определении границ множеств. Попробуем, например, кодировать содержание определенных слов с помощью физических величин, имеющих четкие границы. Так, в сутках 24 часа. Но если мы попытаемся определить с помощью четких величин (часов) границы таких понятий, как «утро», «вечер» и т. п., то возникает определенная нечеткость. С каких часов начинается «утро»? Или определим содержание понятия «юноша». Сколько ему должно быть лет? Такого рода вопросы можно продолжить. Иными словами, многие лингвистические объекты имеют нечеткие, размытые границы.

Лингвисты осознавали нечеткость лингвистических объектов. Ф. де Соссюр, который подчеркивал дискретность этих объектов, признавал в своих работах, что не только у диалектов, но и у языков нет естественных, четких границ. Однако неоднократно выдвигались идеи создания универсальных, четких, дискретных языков, которые семантически были бы сильнее любого естественного [Денисов 1965]. Такие задачи исходили главным образом от математиков и кибернетиков, которые занимались формализацией языка. Семантические и синтаксические модели строились исходя из предположения о том, что лингвистические объекты имеют корпускулярное строение и группируются в четкие, конечные и фиксированные множества. Незжизненность этого постулата обнаружилась только после того, когда теоретико-множественная лингвистика потерпела неудачу в попытках построить на ее основе практические алгоритмы машинного перевода.

Процитируем мнение на этот счет выдающегося ученого Р. Г. Пиотровского, который регулярно сотрудничал с математиками в решении прикладных лингвистических задач.

«Дискретно-логическое мышление человека, которое моделируется естественными языками, в том числе «универсальными» семантическими кодами, языками ЭВМ и т. п., является вторичной суперструктурой, возникающей на основе континуальных потоков сознания, использующих не только сенсорно-моторное восприятие, но и лингвистические механизмы... Указанная суперструктура является всегда хронологически более поздним, а в семантическом отношении более бедным образованием по отношению к континуальному мышлению. Каждый текст, записанный на естественном языке, порождается в результате взаимодействия континуального и дискретно-логического мышления. Автоматическая переработка этого текста, использующая нечеткие толерантные множества, с помощью машинного искусственного языка оперирующего четкими множествами и моделирующего только дискретно-логическую суперструктуру человеческого мышления, всегда связана с огрублением и потерей определенной части смысловой информации, содержащейся в исходном тексте. Отсюда следует, что используемый в лингвистическом автомате четкий и дискретный «универсальный» семантический код принципиально не может быть сильнее в семантическом отношении нечеткого и континуального естественного языка» [Пиотровский 1979].

Возникает вопрос: нужно ли лингвисту заниматься алгоритмами? Пусть он даст описание лингвистических фактов, например, грамматики и лексики, а математик составит алгоритм работы с этими данными, причем в любых целях. Для сравнения рассмотрим пример. Пусть стоит задача сконструировать автомобиль. Что он представляет собой? Это определенное расположение деталей, атомов, молекул, которое можно описать системой уравнений или другим

математическим образом. Однако нет нужды доказывать, что так автомобили никто не производит, поскольку от математического описания расположения деталей до реального воплощения нужно пройти целый ряд промежуточных важных этапов. Человека тоже можно описать математически, как в отдельных функциях, так и целиком. Однако такая модель вряд ли была бы полезной в каких-либо отношениях. Поэтому можно привести такое рассуждение. И автомобиль, и человек есть собрание более мелких блоков, каждый из которых выполняет свою определенную функцию. Эту функцию и нужно описать, и это можно сделать, в том числе и с использованием компьютера. Однако между абстрактными моделями и предметом в целом будут некоторые переходные модели, рабочие модели отдельных блоков, из которых складывается крупная система. Так же можно рассуждать и в отношении лингвистических задач.

Выше было сказано, что одно из направлений в машинном переводе начиналось под лозунгом «Машинный перевод без перевода, без машин, без алгоритмов». В задачу лингвистов предполагалось включить лишь составление описания лингвистических фактов — декларативное описание языка, — а задача составления алгоритмов работы с этим описанием возлагалась на математиков. Если следовать этим рассуждениям, то и, например, задачу обучения языку в виде составления соответствующих алгоритмов, их последовательности тоже можно передать математикам. Однако ясно, что и в первом (перевод) и во втором случае (обучение языку) это сделать невозможно. Нужны постепенный учет и использование человеческого фактора, континуального человеческого мышления. Тем не менее моделировать такой учет и движение в этом направлении безусловно следует, как видно из предыдущих разделов этой главы.

Порождение речи

Рассмотрим структуру языкового знака. Здесь существует несколько концепций. Большинство исследователей придерживаются четырехкомпонентной структуры языкового знака. Ее можно сравнить со структурой знака математического, который вводится в компьютер и в котором некоторые компоненты знака естественного языка не могут быть реализованы. На рис. 13 сопоставлены структуры знака естественного языка и знака математического.



Рис. 13. Структура языкового знака и знака математического

Лингвистический знак (слово, словосочетание, предложение) обычно включает в себя четыре компонента: имя (материальный носитель информации), денотат (отражение предмета из внешнего мира, экстралингвистическая составляющая знака), десигнат (понятие о предмете, отраженное в языке и речи), коннотат (комплекс чувственно-оценочных оттенков, связанных с предметом и понятием о нем). Математический знак имеет только имя и десигнат (другой термин — сигнификат).

Лингвистический знак многозначен. Так, например, слово «стол» в русском языке имеет значение «предмет мебели» и может также использоваться для обозначения меню, состава блюд. Математический знак имеет одно концептуальное значение. Лингвистический знак потенциально метафоричен, у знака математического метафоричность полностью отсутствует. Так, математический знак «семь» имеет только одно десигнативное математическое значение «семь разных предметов», а лингвистический знак «семь» и его производные — «семерка» и пр. — имеют несколько значений — цифра 7, карта в семь очков, группа из семи человек и т. п. При этом это слово несет дополнительные коннотативные значения, связанные с такими словосочетаниями, как «великолепная семерка», «семь смертных грехов», «семь дочерей Атланта» и т. п.

Значение математического знака можно представить в виде множества элементов, имеющего четкие границы — математическое значение — только 7, а не шесть или восемь предметов. Десигнативное значение лингвистического знака также может рассматриваться как множество денотатов, однако оно не имеет четких границ. Например, трудно определить смысловые границы слов «голубой» и «синий», хотя в качестве экстралингвистической реальности здесь можно принять физическую величину длины волн цветового спектра, которые совершенно конкретны. Разные люди могут по-разному называть оттенки этих цветов. В английском и французском языках, например, нет специального слова для прилагательного «голубой».

Значения лингвистических знаков — нечеткие множества с размытыми границами, как определяет этот аспект Р. Г. Пиотровский [Пиотровский 1975]. Особенности построения лингвистического знака приводят к тому, что естественный язык представляет собой нежестко организованную диффузную систему, которая воспринимается и используется человеком в значительной мере интуитивно. Напротив, язык математики является хорошо организованной системой, существующей и функционирующей в виде логического построения, каждый элемент которого имеет осознанную значимость.

Противопоставление языка математики и естественного языка требует, чтобы каждому лингвистическому объекту был поставлен в соответствие математический объект. Лингвистические знаки должны интерпретироваться с помощью знаков математики. Математическая экспликация знаков естественного языка необходима при ре-

шении вопросов, связанных с обработкой текстов этого языка на компьютере.

Р. Г. Пиотровский также предлагает следующую схему порождения текста с учетом основных факторов такого порождения (рис. 14).

Порождение текста определяется, с одной стороны, системой языка и ограничивающей ее действие нормой, а с другой — независимой от языка внешней ситуацией. Рассматривая знак как неколичественную систему, комбинаторная лингвистика описывает механизм перехода от языка к речи с помощью контекстно-свободных грамматик, которые не учитывают контекстные ограничения на употребление отдельных лингвистических единиц и их сочетаний. Эти грамматики порождают много цепочек, не являющихся реальными предложениями данного языка. Чтобы добиться проявления реальных текстов, необходимо перейти к более сильным контекстно-зависимым грамматикам. Такие грамматики можно построить при условии, что к элементам системы языка применяются вероятностные оценки, а сам язык рассматривается как количественная производящая система, функционирование которой регулируется вероятностными ограничениями, заложенными в норме.

Текст представляет собой линейную цепочку отграниченных друг от друга дискретных символов. Каждый из них встречается в тексте с определенной частотой и обладает особыми валентностями, т. е. лингвистическими способностями сочетаться с другими символами. Эти свойства лингвистических единиц эксплицируются в терминах теории вероятностей и математической статистики. Из этого следует, что математическая экспликация центральной проблемы современного языкознания «система языка — норма — текст» может быть осуществлена на основе применения методов как «качественной» (комбинаторной), так и «количественной» математики.

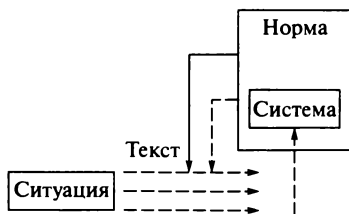


Рис. 14. Схема порождения текста

Общая схема информационного процесса

Исходя из схемы порождения речи можно представить общую схему информационного процесса, т. е. обмена текстами и понимания этих текстов, между отправителем и получателем (рис. 15, с. 106).

В данной схеме тезаурус представляет собой систему понятий, которая содержится в интеллекте как отправителя, так и получателя сообщения. Она оформлена в виде некоторого семантического (смыслового) словаря, в котором каждому слову дается его объяснение в рамках некоторой общей картины мира. Семиозис понимается как пятичленное отношение, включающее имя — носитель информации,

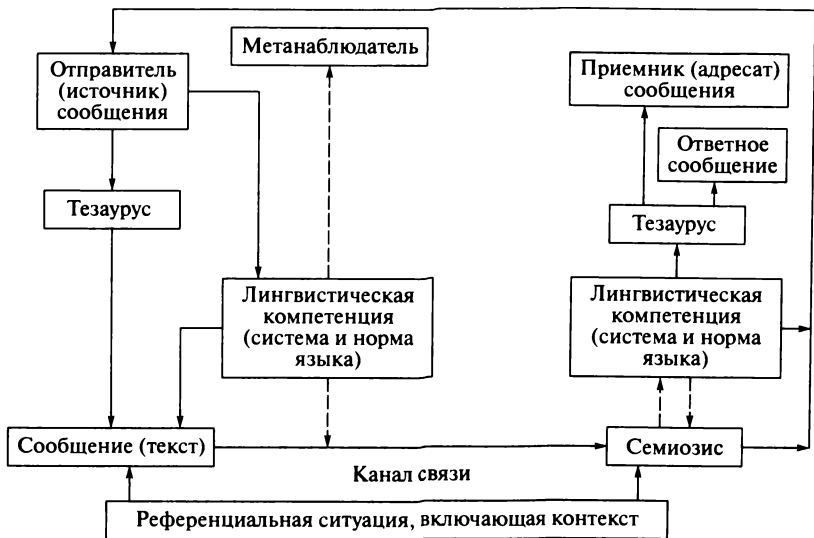


Рис. 15. Общая схема информационного процесса

приемник сообщения — означаемое — ситуационный контекст — интерпретация. В более простых терминах семиозис можно обозначить как понимание, мышление, восприятие сообщения.

Для того чтобы абонент мог принять, расшифровать, интерпретировать речевое сообщение и отреагировать на него, приемник сообщения должен обладать следующими функциями:

- формального отождествления графического, звукового или любого другого выражения сообщения с той формой, которая есть в лингвистической компетенции абонента;
- семантического отождествления лексических и грамматических единиц сообщения с соответствующими единицами, заложенными в его тезаурусе;
- понимания, позволяющей абоненту соотносить содержание сообщения со своим лингвистическим и энциклопедическим (жизненным) опытом;
- ситуативной оценки, определяющей соответствие содержания сообщения с той ситуацией, в которой осуществляется информационный процесс;
- прагматико-рематической, позволяющей выделить в сообщении те элементы, которые являются для получателя наиболее важными и новыми;
- эвристической, позволяющей делать собственные умозаключения и интеллектуальные выводы;
- децизивной, которая позволяет абоненту формулировать решения о дальнейших своих действиях;

– самообучения, дающей возможность приемнику встраивать принятое сообщение или его элементы в свою лингвистическую компетенцию или тезаурус;

– коррективной, которая позволяет абоненту исправлять принятое сообщение в случае его искажения.

Абонент также должен обладать функцией порождения ответа. Этот ответ может иметь либо чисто информирующее задание, либо директивную направленность (приказ, побуждение), либо предусматривать проверку или коррекцию понимания собеседником предыдущей реплики.

Перечисленные выше особенности лингвистического знака вызывают трудности такого моделирования, когда лингвистический знак заменяется знаком искусственным.

Лингвистические аспекты искусственного интеллекта

Интеллект (разум) определяется обычно как способность кибернетической системы — живой или искусственной — целенаправленно решать эвристические задачи, средства, способы и результаты решения которых не заложены в ее прошлом опыте, но выработаны впервые в ходе ее взаимодействия с внешней средой и путем применения прошлого опыта.

Согласно концепции Р. Г. Пиотровского, моделирование разума путем создания систем искусственного интеллекта осуществляется обычно либо в форме математического аналога, либо путем построения технологических устройств, выполняющих операции восприятия, хранения, преобразования и передачи информации, а также осуществляющих на этой основе целесообразное поведение применительно к той или иной окружающей среде.

Общая задача построения систем искусственного интеллекта обычно разбивается на подзадачи, ориентированные на создание программ-блоков, моделирующих прием, хранение, переработку и выдачу информации. Возможны, по крайней мере, три модельных подхода:

– моделирование результатов деятельности разума или отдельных его блоков;

– феноменологический подход, при котором в системе моделируются те приемы переработки информации, которыми пользуется человек;

– структурное моделирование, при котором в системе искусственного интеллекта воспроизводятся те биологические механизмы, с помощью которых мозг перерабатывает поступающую в него информацию.

Общение человека с компьютером и взаимодействие знаков естественного языка и машинных знаков можно представить в виде *двух информационных процессов*.

Первый процесс — общение человека с компьютером и преобразование знаков естественного языка (четырёхкомпонентная структура) в знаки математические. Этот информационный процесс включает четыре блока.

1.1. Порождение текста отправителем сообщения.

1.2. Передача в компьютер цепочки импульсов, кодирующих этот текст.

1.3. Машинный семиозис принятой совокупности импульсов, в результате которого осуществляются распознавание и формализация содержания текста.

1.4. Преобразование в компьютере распознанного текста, например перевод его на другой язык с помощью соответствующих алгоритмов и словарей или компрессия текста с целью индексирования и т. п.

Второй процесс состоит из следующих блоков.

2.1. Выдача компьютером машинных результатов.

2.2. Чтение машинных результатов адресатом.

2.3. Семиозис машинного сообщения.

2.4. Реакция получателя на принятое сообщение.

В обоих случаях успех семиозиса зависит от того, насколько успешно решается вопрос о переходе от естественного языка к языку машины.

Преобразования лингвистического знака в этой системе следующие.

В блоке 1.1 человек порождает лингвистические символы, каждый из которых включает все четыре компонента языкового знака естественного языка.

В блоке 1.2 происходит разрушение каждого из этих символов: путем многократного перекодирования в память компьютера вводится лишь цепочка импульсов, соответствующая означаемому (имени) текстового символа. Денотат, коннотат и десигнат этого символа не попадают в компьютер.

В блоке 1.3 осуществляется машинный семиозис, в результате которого происходит восстановление разрушенного знака. Принятая машиной цепочка импульсов последовательно сравнивается с записанными в машинных словарях цепочками до тех пор, пока не произойдет полное совпадение текстовой цепочки с некоторой словарной цепочкой. При таком совпадении лексико-грамматическая информация словарной цепочки передается текстовой цепочке, которая таким образом получает свой десигнат и превращается в математический искусственный знак.

В блоке 1.4 происходят необходимые действия над знаками текста и осуществляются определенные алгоритмами преобразования.

Второй информационный процесс включает блоки 2.1 — 2.4. В блоках 2.1 и 2.2 происходит новое преобразование знаков текста, которые превращаются в цепочку импульсов. Эти цепочки поступают в блок 2.3, в ходе работы которого человек — получатель сообщения,

опираясь на свои знания предмета и лингвистические навыки, восстанавливает у полученных цепочек их десигнативные и коннотативные значения. Блок 2.4 определяет реакцию абонента на полученное сообщение.

Таким образом, при общении человека с компьютером происходит преобразование знаков естественного языка в знаки языка искусственного и затем восстановление естественно-языковых знаков в процессе семиозиса человека. Достаточно полное восстановление информации может произойти при соблюдении следующих условий:

- должно перерабатываться сообщение, основное содержание которого не коннотативное (мимика, жесты), а десигнативное и со-держится в десигнатах лексических единиц;
- тексты должны быть ориентированы на приблизительно одинаковые тезаурусы внешнего мира;
- обучение компьютеров должно быть ориентировано на однородные тексты, приблизительно равно соотносящиеся с тезаурусом;
- вводимое в компьютер описание языка должно содержать такие лексико-грамматические сведения, использование которых свело бы к минимуму потерю информации при переходе от символов входного сообщения к текстовой последовательности машинных знаков, при преобразовании машинного текста в компьютере и при расшифровке приемником сообщения машинных результатов.

Лексико-грамматический анализ текста в формальном аспекте

Лексико-грамматический анализ является важной частью автоматического анализа текстов на естественных языках, который в настоящее время распространен в системах автоматической обработки текстов. В мире существует огромное количество программного обеспечения для поиска, хранения и обработки информации — программы автоматического и синтаксического анализа текстов, аннотирования и реферирования документов, системы машинного перевода, различного вида пользовательские интерфейсы на естественном языке, автоматические интерактивные словари, экспертные системы и т. п. Однако в общем программы обработки естественного языка далеки от совершенства.

Сейчас существует очень много различных теорий автоматической обработки естественного языка, которые применяются в системах машинного перевода Pars, Prompt, Stylus, Socrat, Word Translator и др., в системах реферирования и аннотирования Libretto, AutoSummarize, Intelligent Text Miner, МультиЛекс, в системах поиска и классификации информации Google, AltaVista, Yandex, Goldfire Intelligence, TREC и во многих других. В данном случае мы пользуемся обзором и ре-

зультатами аналитической обработки информации, которые содержатся в работе С. Ю. Кравченко [Кравченко 2008].

Задача любой системы обработки естественного языка — «расшифровать» смысл высказывания на естественном языке и представить его в формальном виде, понятном компьютеру и доступном для дальнейшего использования. Особенно ценным и необходимым решением этой задачи представляется для машинного перевода, который в современном состоянии не является полностью удовлетворительным по качеству. «Благодаря формализации станет возможным сравнение грамматик самых разных по строению языков и выявление стоящих за ними универсальных механизмов, допускающих математическое исследование» [М. Гросс, цит. по: Кравченко 2008]. Исследователями было разработано множество теорий формализации естественного языка, но на практике, при осуществлении их реализации в виде действующих систем, возникало и возникает много трудностей.

Методы, которые применяются для формального разрешения этих проблем, С. Ю. Кравченко группирует следующим образом:

- вероятностно-статистические методы;
- методы, основанные на правилах;
- полностью автоматические методы, или использующие «нейронные сети»;
- гибридные, или комбинированные, методы, использующие одновременно несколько из перечисленных методов.

Каждый из этих методов имеет свои достоинства и недостатки, каждый может быть применен к различным языкам с разной степенью эффективности.

Вероятностно-статистические методы. В настоящее время, по наблюдению С. Ю. Кравченко, это один из наиболее эффективных методов лексико-грамматического анализа. Простейшим примером его работы является решение, когда словам приписываются лексико-грамматические классы с опорой на вероятность, с которой данное слово принадлежит именно этому классу. Например, если какое-нибудь слово чаще всего употребляется в роли глагола, то оно и будет глаголом в большинстве случаев.

Методы, основанные на правилах. Этот подход также активно используется при анализе. Суть его заключается в том, что для определения лексико-грамматического класса слова могут использоваться наборы специальных правил — контекстных правил или шаблонов. Они описывают контекст, по которому можно однозначно отнести слово к тому или иному классу или, по крайней мере, ограничить неоднозначность. Шаблоны могут создаваться или автоматически, или человеком, который на основе исходного конкорданса может создавать правила разрешения омонимии для конкретных слов.

Пример контекстного правила: если многозначному слову предшествует определитель, например, артикль, а само оно предшествует существительному, то оно есть прилагательное. По такому правилу,

например, в предложении *A fat horse is climbing up the hill* слово *fat*, которое может быть существительным или прилагательным, определяется как прилагательное.

Пример ограничения многозначности (из этапа разрешения лексико-грамматической омонимии системы машинного перевода АМПАР): английская словоформа *round* может выступать как существительное, прилагательное или глагол. При наличии окончания *-s* — *rounds* — оно уже не может быть прилагательным, хотя выбор между существительным и глаголом остается. Далее в алгоритме разрешения такой неоднозначности должны быть предусмотрены и другие проверки контекста.

Использование метода, основанного на правилах, или, точнее, на контекстном анализе лингвистического окружения омонима, получило большое распространение в системах автоматической обработки.

Автоматические методы. Этот подход является почти полностью автоматизированным и не требует данных, предварительно подготовленных человеком, — специальных словарей, правил анализа контекстов и т. п. Используя довольно сложные вычисления, соответствующие алгоритмы автоматически порождают контекстные правила и частотные словари. Центр разработки при этом ложится на статистические методы [Пиотровский 2006].

Комбинированные (гибридные) методы применяют одновременно несколько вышеперечисленных подходов для автоматического лексикографического анализа текстов. Очень часто используется комбинация первого и второго подходов. Например, после анализа текста с применением статистической информации полученные результаты уточняются с применением правил анализа контекста. Исследования показывают, что использование нескольких типов анализа позволяет значительно повысить точность лексико-грамматического разбора текстов.

Все перечисленные методы активно используются в современных системах обработки естественного языка. Выбор того или иного подхода обусловлен, прежде всего, типом естественного языка, с которым работает система. Например, для обработки языков аналитического типа (английский, французский, испанский, итальянский и др.) обычно применяются вероятностно-статистические методы анализа с последующим уточнением результатов при помощи контекстных правил. Такой выбор обусловлен тем, что языкам аналитического типа свойственна синтаксическая многозначность слов (конверсионная омонимия), что затрудняет анализ текста с помощью только контекстуальных правил. Например, было бы проблематично распознать правильные классы слов в английском предложении *The brown dog bites* — *Коричневая собака кусает* или *Укусы коричневой собаки* только с помощью анализа контекста. В данном случае можно привлечь статистические оценки, которые могут дать наиболее правдоподобный вариант разбора.

Для обработки флективных языков (русского, белорусского, чешского и др.) используют в основном правила морфологического анализа или контекстные правила. Это связано с большой формальной выраженностью грамматических характеристик слов в их текстовых словоформах.

При обработке текстов на тайском, китайском, японском и других иероглифических языках, где основную трудность представляют иностранные или неизвестные слова и нехватка языкового материала для сбора статистической информации, наиболее подходящим вариантом является комбинация автоматических методов, морфологических и контекстных правил.

Учитывая все вышесказанное, заключает С. Ю. Кравченко, возникает закономерный вопрос — каким должен быть лексико-грамматический анализ, например, в системе машинного перевода, работающей с английским, французским, немецким и русским языками? Наиболее подходящим, по его мнению, является следующий вариант:

- морфологический разбор по формальным признакам с целью установления лексико-грамматических классов слов;
- статистический анализ для выбора наиболее вероятной последовательности слов с определенными лексико-грамматическими классами;
- корректировка результатов статистического анализа с помощью контекстных правил, созданных человеком.

Можно, однако, заметить, что первый способ, т. е. морфологический разбор по формальным признакам, также производится с помощью алгоритмов, разработанных человеком на основе предварительно полученного словаря-конкорданса с помощью методов корпусной лингвистики.

Контрольные вопросы

1. Какие технические и научные достижения обусловили появление идеи машинного перевода?
2. Какие основные этапы работ по реализации идеи машинного перевода вам известны?
3. Существует ли системность в лексике естественного языка?
4. Какое содержание придается в современном языкознании понятию «подъязык»?
5. Какова роль количественной лингвистики в формализации процесса перевода?
6. Влияет ли типология текстов на эффективность машинного перевода и в какой степени?
7. Каковы современное состояние и практическая роль действующих систем машинного перевода?

РАЗДЕЛ III

МАШИННЫЙ ПЕРЕВОД КАК ЦЕНТРАЛЬНАЯ ПРОБЛЕМА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Глава 10

ЗНАЧЕНИЕ ИДЕИ МАШИННОГО ПЕРЕВОДА

Машинный перевод и теория языка

Идея машинного перевода, т. е. мысль поручить машине работу по переводу с одного естественного языка на другой, насчитывает к настоящему времени уже около пятидесяти лет существования. Примерно столько же ведутся научно-исследовательские и опытно-конструкторские работы по машинному (автоматическому) переводу во многих странах мира. Параллельно с созданием систем машинного перевода разрабатывались и разрабатываются теории, которые можно было бы положить в основу создания таких систем. Интересно отметить, что такие теории вследствие сложности задачи, о чем мы будем говорить далее, как правило, охватывают не только в собственном смысле перевод с одного естественного языка на другой, но и многие смежные актуальные и часто глобальные теоретические вопросы человеческого владения и пользования языком.

Интенсивность работ по машинному переводу в разные периоды времени была различной. Нынешний этап характеризуется некоторым спадом интереса к проблеме по сравнению, например, с 60-ми гг. Мода на машинный перевод прошла, многие его энтузиасты давно занялись другими делами. Современное состояние машинного перевода, если его характеризовать несколько критически и в сопоставлении с тем, что имело место ранее, характеризуется некоторыми вялотекущими научными исследованиями, существованием определенного количества практически работающих и продающихся на рынке программного обеспечения систем машинного перевода в разных странах и появляющимися время от времени сообщениями о разработке новых систем.

Актуальность автоматизации перевода продолжает оставаться высокой вследствие того что перевод с одного языка на другой — единственный эффективный способ преодоления языковых барьеров,

поскольку другие (внедрение и использование универсального языка типа эсперанто, изучение иностранных языков основным составом коммуникантов и пр.) неэффективны и не могут заменить перевода; растут и расширяются возможности современной компьютерной информационной технологии, поэтому всегда появляется соблазн поручить машине какую-нибудь интересную интеллектуальную задачу; спрос на переводы в мире увеличивается в абсолютных и относительных пропорциях соответственно тому, как все больше естественных языков приобщаются к мировой цивилизации и вступают в коммуникационную информационную сферу. Высока также и научная привлекательность проблемы машинного перевода.

Чем интересен машинный перевод для теории языка?

Если вспомнить появление самой идеи машинного перевода и связанных с этим сенсаций, когда компьютеры, только что появившиеся в научной практике, были окутаны ореолом тайны и с ними общались только посвященные, то приходится признать, что эта идея произвела огромное влияние на лингвистику. Появилось множество всяких новых лингвистик: математическая, статистическая, алгоритмическая, вычислительная, инженерная и др. Все они отражали новое направление в языке, новые взгляды на объект, предмет, методы и результаты теоретических изысканий в области естественного языка. Возникли предположения о возможности математического моделирования языка, стал оспариваться тезис о том, что язык внеположен мере и числу, ученые вернулись к тем формальным поискам, точнее, поискам формализмов в языке, которые были известны из трудов выдающихся лингвистов, таких, как Л. В. Щерба, который своей «глокой куздрой» обратил внимание на значение формы в языке, до того находившейся в некотором загоне из-за преобладания семантических взглядов.

Основной смысл новых лингвистик заключался в попытках формализовать разные языковые уровни. Привлекательность машинного перевода для теории и заключалась именно в том, что все эти уровни были разные: задача машинного перевода требует распознавания графических образов (на этапе ввода в компьютер), морфологического анализа, анализа и перевода лексики, слов и словосочетаний, синтаксического анализа и синтеза и, наконец, семантических преобразований, долженствующих обеспечить смысловое равенство введенного и выведенного предложения и/или текста в целом.

Полученная в результате анализа информация должна быть необходимой и достаточной для синтеза переводящих единиц на каждом из уровней.

В этом, т. е. во всеохватываемости, включенности всех уровней языка в задачу, и заключается, по-видимому, **первая** теоретически привлекательная **особенность** машинного перевода.

Вторая особенность состоит в том, что перевод — единая задача, пронизывающая все уровни языковой структуры. Анализ и син-

тез единиц всех уровней служат одной конкретной цели — передаче смысла, содержания высказывания (в широком смысле) с одного естественного языка на другой. Разные языковые уровни можно изучать с разными целями. При этом предмет исследования часто становится неопределенным. Теоретикам языкознания и преподавателям лингвистики хорошо известно, что единого курса языкознания нет и быть не может. Каждый профессор читает свой собственный курс языкознания, исходя из собственного понимания основополагающих сущностей этого предмета. Если математика, химия, физика, геометрия и другие точные науки исходят из нескольких фундаментальных теорем или посылок, на основе которых возникают все дальнейшие построения, то в языкознании все зависит от научной школы, в рамках которой определяются исходные понятия: что принимается за фонему, морфему, как определяется слово или словосочетание, исходит ли автор из концепции частей речи или дистрибутивных классов слов, являются ли объектом лингвистики (по мнению данного ученого) только предложения и фраза или же связный текст целиком и т. п. и т. д. «Внеположенность» языка мере и числу затрудняет квалификацию объектов языкознания в терминах точных наук, что и приводит к множественности теорий языка. Задача же формализации и алгоритмизации для перевода заставляет выбирать некоторые сквозные единицы описания уровней языка, такие, на которых можно было бы строить не только алгоритмы анализа и синтеза отдельных уровней, но и всю систему, с расчетом на соответствующий синтез выходного предложения и текста в целом. Определяющим становится некоторый функциональный подход, налагающий ограничения на безграничные фантазии и «полет творческой мысли», не подкрепленный обратной связью инженерных проверок.

Для добросовестного исследователя в машинном переводе всегда открыта возможность проверить правильность теорий и концепций посредством практики. Сделал алгоритм синтаксического анализа — проверь на материале новых предложений, не учтенных при его составлении, как данный алгоритм работает, какие результаты выдает. Сама задача машинного перевода, а именно получение выходного предложения, соответствующего входному, есть уже реализация принципа обратной связи. Невозможность проверить правильность заложенной концепции можно, видимо, считать признаком теоретического или практического дефекта системы. Если в начале работ по машинному переводу многие чисто теоретические концепции нельзя было проверить, по мнению их авторов, вследствие недостаточной памяти или быстроедействия тогдашних ЭВМ, то теперь такой аргумент отпадает вследствие того, что современные компьютеры имеют значительные объемы памяти и быстроедействия, практически достаточные для любой объемной лингвистической задачи.

Третья особенность машинного перевода — это его связь с лексическим уровнем языка. Примитивный взгляд на перевод — а ведь

именно он лежал в основе самой идеи машинного перевода, выдвинутой, как известно, математиками и инженерами-программистами, — состоит в том, что нужно переводить слова и словосочетания. Такое видение является следствием представления о языке как о некотором коде в прямом, криптографическом смысле слова. Довольно скоро представление о языке как коде было отвергнуто как недостаточное для описания природы языка, но интерес к лексическому уровню, что весьма естественно, остался. Все-таки значительная часть содержания высказывания заключается именно в словах, а не только и не столько в их порядке и согласовании [Комлев 1992]. Без перевода слова и его окружения не может быть проведен анализ предложения, без исследования поведения слова в тексте не может быть вообще никакого перевода, не только машинного. Лексический уровень имеет особое отношение к семантике, а передачу смысла мы на интуитивном уровне считаем главной задачей перевода. Как бы то ни было, лексика для машинного перевода играет большую роль, и именно она объясняет разного рода ответвления от него в виде автоматических словарей в помощь переводчику, терминологических банков данных и пр.

Четвертая особенность машинного перевода — на другом конце лингвистического спектра. Операция перевода является сложнейшей задачей интеллекта, поскольку требует не только замены слов одного языка словами другого, но и передачи мысли в полном ее объеме, со всеми ее оттенками и коннотациями, будь то политический, научный или технический текст. О машинном переводе художественного текста говорили только крайние энтузиасты. Тем не менее и в более точных типах текста и стилях, вроде научного и технического, существует немало количество так называемых переводческих трудностей, связанных с формулировкой мысли в двуязычной ситуации. Как ни определять мысль, мышление, понимание, восприятие смысла и содержания, ясно, что машинный перевод, поскольку он «перевод», должен передать это содержание на уровне текстовых единиц. Отсюда следует, что задача машинного перевода есть задача искусственного интеллекта, какое бы содержание ни вкладывалось в понятие «искусственный интеллект».

Эти четыре особенности задачи машинного перевода как моделирования процесса перевода для компьютера представляли ранее и представляют сейчас фундамент, на котором базируется интерес к нему, и этот интерес не является преходящим (см. также многие современные работы на эту тему, например [Miram 1998, Tsujii 1997] и др.).

Почему эту проблему можно считать центральной в искусственном интеллекте?

По нашему мнению, центральным в искусственном интеллекте является вопрос моделирования деятельности мозга как орудия мышления. Мышление неразрывно связано с языком. Эту связь никто

никогда не отрицал независимо от того, какие роли отводились каждому компоненту. Язык — непосредственная действительность мысли. Моделирование языковой деятельности на уровне перевода затрагивает, как показано выше, все уровни языковой структуры, от морфологии до семантики. Поэтому задача машинного перевода — моделирования сложнейшего комплексного вида языковой деятельности — является первостепенной для искусственного интеллекта. Добавим сюда и то обстоятельство, что машинный перевод как воспроизводящая инженерно-лингвистическая модель обладает таким компонентом, как обратная связь, что позволяет оперативно корректировать исходную систему гипотез [*Пиотровский* 1975, 1979 и др.].

К истории машинного перевода

История машинного перевода чрезвычайно интересна и поучительна. Он начался с появления компьютеров и с попыток применения их к решению интеллектуальных задач. Первыми специалистами по нему были ученые разных отраслей, среди них главным образом математики, программисты, теоретики той области знаний, которая впоследствии стала называться искусственным интеллектом. Первоначальной точкой зрения на машинный перевод была концепция языка как некоторого кода, и проблема перевода приравнивалась к задаче перекодирования из одного языка (одного кода) в другой. Однако достаточно скоро стала ясна односторонность такого мнения, после чего начались углубленные изыскания в теории языка, в попытках главным образом найти способы передачи смысла.

История ранних исследований в области машинного перевода хорошо отражена в аналитическом труде и сборнике воспоминаний ветеранов машинного перевода, выпущенном Дж. Хатчинсом в 2000 г. [*Early Years in Machine Translation* 2000]. Он включил в сборник всю историю машинного перевода, начиная с изобретения русского ученого П. П. Смирнова-Троянского, который предложил механического переводчика, автоматически подбирающего словесные эквиваленты для единиц входного языка, до широкомасштабных исследований, которые во второй половине XX в. развернулись практически во всех странах и привели к возникновению компьютерной лингвистики, а также дали значительный стимул к развитию общей лингвистической теории.

В советской лингвистической традиции исследования по машинному переводу периода 1949 — 1963 гг. отражены в библиографии [*Мельчук* и др. 1967].

Современный исследователь найдет в истории машинного перевода много интересных идей, некоторые из которых просто не могли быть осуществлены в то время из-за слабых возможностей компью-

теров, а другие показали свою практическую неэффективность в том, что касается собственно перевода, но могут представлять интерес с других точек зрения и в свете новых возможностей информационных технологий.

Как и всякая история, без изучения того, что было, трудно ожидать успеха в новых разработках.

Современное состояние машинного перевода

Что же мы можем видеть сейчас, через более чем пятьдесят лет с того момента, когда на машине IBM 701 русский текст «Величина угла определяется отношением дуги к радиусу» был переведен на английский язык без участия человека, чем завершился так называемый Джорджтаунский эксперимент, начавший эпоху машинного перевода? Мы видим довольно пеструю и противоречивую картину.

С одной стороны, машинный перевод есть. В России (в Москве) можно сейчас купить приличную систему машинного перевода с русского языка на английский (и с английского на русский) под названием «Сократ» за 149 долларов, а также системы машинного перевода «Стилус», «Промпт» и др. Система фирмы «Бит» стоила 300 долларов. В США такие продукты стоят от 1 000 долларов и качество перевода не лучше, чем дают наши системы. Эти затраты не совместимы с расходами на разработку систем машинного перевода «с нуля». Особенно дорого и сложно разработать эффективное программное обеспечение. Демонстрация работы систем показывает, что перевод получается удовлетворительный. Даже тексты, принесенные потенциальным заказчиком и никак не учтенные при разработке и составлении словарей, иногда получают приличный вид. Однако в большинстве случаев перевод требует пост-редактирования. Работать с современной системой машинного перевода переводчику, не знакомому или плохо знакомому с программированием и с работой на компьютере, довольно сложно. Системы, как правило, обладают защитой от переписывания, вносить изменения в алгоритмы и программы нельзя. Словари пополнять и изменять можно, однако часто лишь с помощью специальных программ, продающихся отдельно. Ввод текста происходит через сканер, с ручным исправлением искажений. Эта часть тоже продается отдельно. Системы допускают пост-редактирование.

Несмотря на невысокую стоимость, приемлемое качество и потребность в переводах, объем продаж, как кажется, невелик, и, насколько известно, не так много фирм пользуются машинным переводом.

Такое положение можно объяснить следующим:

– переводом в фирме занимается переводчик, он системным программистом не является, и свободное обращение с программой пе-

ревода, насколько бы дружественной к пользователю она ни была, для него дело довольно сложное. Прикомандировывать программиста — возрастают расходы. Редактировать плохой машинный перевод занятие трудоемкое и весьма неохотно выполняемое редакторами и переводчиками, которым часто бывает проще перевести текст самим «с нуля», чем выискивать и исправлять ошибки;

– переход на другие тематики и другие типы текстов, что весьма распространено в переводческой деятельности и осуществляется человеком без особого труда, в машинном варианте всегда довольно затруднителен и может быть вообще невозможным при больших различиях в текстах.

Учитывая невысокое качество перевода и высокую стоимость обслуживания, другие неблагоприятные факторы нынешней информационной ситуации, применение современных действующих систем машинного перевода не является распространенным, и их вклад в преодоление языковых барьеров (хотя точные цифры и отсутствуют), видимо, невелик.

Тем не менее, принимая во внимание общую ситуацию с переводами в мире, можно смело утверждать, что альтернативы машинному переводу нет. Какие соображения говорят в пользу такого суждения?

Основные проблемы машинного перевода

Представляется, что оптимальное решение проблемы машинного перевода, точнее, наиболее актуальное приложение усилий к решению этой проблемы, может быть достигнуто только в том случае, если мы определим приоритеты научных исследований в этой области.

В этой связи возникают два основных направления:

- попытка сделать так, чтобы переводился смысл безотносительно к форме языкового высказывания;
- осуществление машинного перевода на уровне переводных соответствий, когда передаются в равной степени и форма, и содержание.

Казалось бы, эти два направления не противоречат, а как бы дополняют друг друга. Однако на самом деле в истории машинного перевода они постоянно находились в противоборстве, что в значительной мере мешало эффективному развитию исследований и разработок.

Первое направление можно обозначить как «смысл-текст», или «текст-смысл», а более точно, применительно к собственно переводу, «текст-смысл-текст». Исследования в этом направлении имели целью разработать универсальный смысловой язык-посредник, а когда стала ясна недостижимость этой цели (довольно быстро, впрочем), стремились создать отдельный семантический компонент систем,

чтобы он контролировал и переводил семантику как высший уровень языковой структуры в переводе. Эти попытки также практически провалились, ни одной действующей системы с семантическим компонентом мы до сих пор не имеем нигде в мире, хотя работы велись широким фронтом во многих странах, в том числе и в СССР.

Второе направление можно назвать моделированием перевода на уровне переводных соответствий. Оно возникло из начальных воззрений инициаторов машинного перевода относительно языка как разновидности кода [Марчук 1983]. Кроме того, в машинном переводе решающая роль отводилась словарю. Сторонники передачи смысла иногда именовали данный подход «методом грубой силы» (*brute force approach*). Однако игнорирование этого подхода приводило вообще к отказу от машинного перевода. Что значит «передать смысл» в переводе? Смысл предложения, высказывания, абзаца и целого текста можно без особых трудностей выразить либо одним словом, либо парой слов, либо цифрой, кодирующей данный «смысл» в какой-то знаковой системе (например, информационного поиска). Но ведь в переводе должен быть передан не такой «смысл», а смысл на уровне и формы, и содержания. Поэтому отказ от учета именно переводных соответствий и означал отказ от перевода вообще.

Тем не менее долгое время и ценой многих усилий исследования велись в этом тупиковом направлении. Тупиковость его стала особенно очевидной именно сейчас.

В отличие от этого моделирование перевода на уровне переводных соответствий и разные модификации этого подхода (иконический перевод и т. п.) дали реально работающие системы и позволили осмыслить как теорию машинного перевода, так и практику, намечая эффективные пути совершенствования (см. также [Пиотровский 1999]).

Проблемы современного машинного перевода

Основные проблемы современного машинного перевода удобно рассмотреть на материале проблемного доклада известного французского специалиста К. Буатэ, одного из сторонников использования смысла в машинном переводе в виде языка-посредника. Одновременно в течение долгого времени он занимался созданием систем, которые могли бы реально переводить тексты. В своем докладе на международной конференции по актуальным проблемам вычислительной лингвистики в Пенанге, Малайзия, он сформулировал двенадцать проблем современного машинного перевода [Boitet 1991]. К. Буатэ выделяет три вида современного машинного перевода:

– **«информативный»** перевод. Это грубый машинный перевод, пословный, достаточный для поверхностного знакомства с содержанием текста на незнакомом языке;

– **«профессиональный»** машинный перевод. Его качество сравнимо с качеством «человеческого» перевода и при небольшом редактировании является полностью удовлетворительным. Выгоден для больших объемов текста (свыше 10 000 страниц в год) и для однородных текстов;

– **«персональный»** перевод. Авторы входных текстов заранее избавляют их от неоднозначностей и работают в режиме диалога с компьютером. Возникает возможность существенно улучшить качество перевода.

Двенадцать проблем современного машинного перевода К. Буатэ делит на четыре класса: концептуальные, инженерные, технические и проблемы архитектуры. Имеет ли все это отношение к лингвистике? Прежде чем отвечать на этот вопрос, удостоверимся, что согласно принципам «нового структурализма» (да и старого здравого смысла) к объектам лингвистики принадлежат объекты, изучаемые лингвистом-практиком, которым лингвист-теоретик может дать теоретическую оценку. Эти объекты не могут устанавливаться независимо от лингвистической практики [Prospects 1992]. **Концептуальные проблемы** формулируются К. Буатэ следующим образом:

– каким образом представлять и разрешать неоднозначности всякого рода? Каждой единице перевода сопоставляется, как правило, несколько возможных абстрактных представлений. Надо не только выбрать нужное в данном тексте, но и предвидеть появление других неоднозначностей в результате данного выбора. Для решения этого вопроса целесообразно ввести «неоднозначное программирование», предусматривающее разные возможности разрешения неоднозначностей;

– как обеспечить точное переводное соответствие между абстрактной структурой и текстом? Конкретное высказывание всегда отличается от его абстрактной модели. Концептуальной проблемой является также использование формализмов. Каким образом создать эффективные средства анализа из чистых формализмов, добавляя к ним процедурные или эвристические знания? Ни одна из действующих систем машинного перевода не использует только декларативные знания.

Если из этого текста вычленим «неоднозначное программирование», смысл которого плохо понятен лингвисту и которое является не более чем некоторым спасательным кругом, помогающим математику и программисту, коим является К. Буатэ, не утонуть в проблеме, то получаем следующее:

– каждая единица перевода может иметь несколько абстрактных представлений;

– необходимо предвидеть появление других неоднозначностей при движении далее по тексту от найденной, а также учесть в максимальной степени все другие возможные неоднозначности, могущие возникнуть в результате принятого решения;

- у текста (не только у единицы перевода) может быть несколько абстрактных структур, и неизвестно, как найти нужную;
- чистый формализм не может служить основой анализа;
- к формализму должны быть присоединены декларативные или процедурные лингвистические знания;
- способ присоединения таких знаний неясен.

О переводе смысла

Пессимист, знакомый с историей машинного перевода, может сказать, что эти проблемы были ясны, по крайней мере, тридцать лет тому назад и что мы пришли опять к тому, от чего, вроде бы, должны были продвинуться. Однако оптимист, также знакомый с историей машинного перевода, заметит, что изложенные здесь проблемы сформулированы исследователем, всю свою деятельность посвятившим машинному переводу с помощью языка-посредника. В рамках этого подхода предполагалось перевести входной текст сначала на язык-посредник, а с него на любой другой естественный язык — выходной. При таком осуществлении перевода все названные выше проблемы оказались бы за скобками — причем тут какие-то неоднозначности, если переводится смысл, а само понятие смысла исключает неоднозначности на всех предшествующих уровнях. Если мы знаем смысл высказывания, то все неоднозначности уже разрешены. В чем же тут дело?

Рассмотрим сначала концепцию языка-посредника. Известно, что под таким языком понималось многое, от языка семаглицов до пучков грамматических соответствий. Однако никто, включая авторов проектов, не знал, как их строить и как практически применять к переводу [*Мельчук* и др. 1967]. Дело, однако, было не в этом, а в обосновании концепции «машинный перевод без перевода, без машин, без алгоритмов». Этот трюк и стремление к мировой революции в лингвистике на волне энтузиазма по поводу возможностей компьютера дорого обошлись науке и значительно задержали исследования и разработки практически ориентированных систем. Многие подобные абстракции давно и благополучно забыты. Рассматривая вышеизложенные постулаты К. Буатэ, важно учесть прежде всего то, что они сделаны единственным последовательным сторонником идеи машинного перевода через язык-посредник, функцией которого является использование некоторого смыслового языка.

Трудности перехода на язык смысла от нормального естественного языка чрезвычайно велики. Десятилетия работ по машинному переводу показали, что обращение к смыслу переводимого, такое вроде бы простое и естественное для человека-переводчика, в машинном моделировании чрезвычайно сложно. Начать с того, что понятие «смысл» трудно вразумительно определить. В. А. Звегинцев убедительно показал, что предложение, взятое вне контекста, имеет не

смысл, а «псевдосмысл», а разные перефразы одного и того же высказывания имеют и разные смыслы. Он приводит перефразы простого предложения: «Охотник изо всей силы ударил волка ногой» — «Волк получил удар ногой от охотника. Волк был повергнут ударом ноги существа мужского пола. Внезапный удар ногой сразил волка» и т. д. Содержание действий, соотношение субъекта и объекта вроде бы одно и то же, но одинаков ли языковой смысл вариантов высказывания? Без особых мудрствований видно, что нет [Звегинцев 1976]. А что такое смысл текста? Это сумма смыслов отдельных высказываний? Если сумма, то какая — арифметическая или некоторая интегральная, зависящая от ситуативного и коммуникативного контекстов или нет? Список вопросов, на которые нет вразумительных и конструктивных с лингвистической точки зрения ответов, можно продолжить. Характеристикой положения является тот факт, что до сего времени нет ни одной действующей системы машинного перевода, которая в явном виде использовала бы понятие смысла и / или семантического компонента в объеме, большем чем для двух-трех примеров.

Что все это означает с точки зрения теории машинного перевода? Прежде всего то, что победила теория и концепция «текст-текст», а не «текст-смысл-текст». Действующие системы машинного перевода так или иначе реализуют принцип перевода с помощью переводных соответствий, без всяких семантических заглублений, больших, чем необходимо для устранения чисто переводных многозначностей.

Другие проблемы машинного перевода

Из вышесказанного следует, что теория машинного перевода должна развиваться в сторону дальнейшего осмысления эквивалентных, вариантных и трансформационных соответствий в рамках данной языковой пары с последующим переходом (когда нужен множественный перевод) к другой языковой паре. Неслучайно К. Буатэ особо выделяет переводные соответствия, о которых он ранее не говорил. Думается, что теория должна выявить природу возникновения, характер, распространенность, возможности формализации таких соответствий, структуру и объем процедурных знаний, приписываемых каждому типу соответствий.

Рассмотрим теперь **проблемы «архитектуры»** в концепции К. Буатэ. Первой из таких проблем является соотношение между вариантами и настройкой. Вопрос ставится так: как конструировать и использовать грамматические формализмы, которые позволили бы описать варианты (микроязыки или подязыки) в рамках единого языка, приспособив конкретную систему перевода для данного естественного языка к конкретному подязыку посредством настройки?

Здесь прежде всего выделяется лексический аспект. Невозможно содержать огромные базы лексических данных, которые были бы

пригодны для каждого подязыка в рамках данного языка. Совершенно также не исследован грамматический аспект (имеет ли каждый подязык свою грамматику или же пользуется единой грамматикой языка?). Для решения этой проблемы К. Буатэ предлагает объединение статистических и содержательных подходов. Так, статистика грамматических конструкций в рамках данного подязыка может быть получена из текстов. Содержательный анализ позволит выявить наиболее актуальные типы конструкций. Тогда подязык мог бы быть определен как совокупность конечного инвентаря лексики и набора грамматических правил. Вопрос этот также не нов. Практика действующих систем машинного перевода пока еще не дала ответа на вопрос о том, как лучше организовать словари и грамматику в условиях разных подязыков одного и того же языка. Можно лишь согласиться с К. Буатэ в том, что лингвистическая теория также еще не предложила каких-либо решений этой проблемы. Когда они будут найдены, то еще предстоит научить пользователя выбирать тот или иной подязык.

Обучение на основе текстов. Возможно ли создать системы машинного перевода, которые будут реально учиться на основе существующих параллельных текстов и на корректировке ошибок, отмеченных профессиональными редакторами?

Вопрос об обучении систем машинного перевода на основе параллельных текстов также не нов, однако реально мало что достигнуто на этом пути. Представляется интересным повторное обращение к статистическим методам в построении таких систем, в том числе и с интерлингвой (см., например, последние японские работы, доложенные на Московской международной конференции по квантитативной лингвистике в 1994 г. [Tsutsumi 1994]). К. Буатэ также положительно оценивает роль параллельных текстов при составлении как словаря, так и грамматики. Остается лишь напомнить, что параллельные тексты служили основой и были главными критериями при разработке словарей и алгоритмов системы АМПАР еще в 60-х гг. [Марчук 1970; Моторин 1970].

Степень понимания. Как воспроизвести «нечеткое» понимание человека-переводчика? Как воспроизвести высказывания частично в content-bound way, а частично в structure-bound way? Цитируем К. Буатэ: «Основанный на знаниях машинный перевод возможен в некоторых контекстах, где можно положиться на полное знание и полное представление некоторой предметной области или задачи. Например, можно переводить с помощью машинного перевода инструкции по уходу за ядерными реакторами. Но если пытаться распространить его на все области, то это невозможно». Нетрудно увидеть, что это означает отказ от передачи смысла. Ведь в ограниченной предметной области, например в инструкциях, весь перевод реально осуществляется в рамках переводных соответствий и никакого обращения к смыслу (content) вообще не требуется.

Связывание архитектуры с применениями. Как увязывать архитектуру с типами применения? Здесь и далее излагаем концепцию К. Буатэ. Помимо своей существующей теоретической значимости машинный перевод представляет собой инженерную проблему. Предложены многие архитектуры для разных применений (трансфер-интерлингва, пакетный-интерактивный режим и т. п.), не говоря о разнообразии лингвистических теорий и компьютерных инструментов. Но ни одна из них не годится для всех применений сразу.

Разработчики имеют свои идеи относительно адекватности архитектур и их применений. Проблема здесь заключается в согласовании действий разработчиков и заказчиков. Заказчик часто имеет свои собственные предпочтения, основанные на интуиции. Его нельзя переубедить в том, что такой-то путь неправилен, не пытаясь его реализовать. Но потом он теряет вместе с большими деньгами и весь интерес к машинному переводу.

Поэтому важная проблема — дать четкую картину того, что можно ожидать и чего нельзя ожидать от машинного перевода.

Измерение прогресса в машинном переводе. Как измерить прогресс и вклад исследований по переводу в фундаментальные представления о характере языка и его использовании?

Наподобие использования термоядерной реакции, машинный перевод является мечтой нашего века. В противоположность ей он уже дает результаты. Однако финансируется далеко не так, как термоядерные исследования. Почему? Возможно, одной из причин является то, что отсутствует чисто научный элемент, а именно измерение прогресса. Машинный перевод и связанные с ним исследования, видимо, не вносят какой-либо измеримый, фундаментальный вклад в науку.

Даже меньшую цель, а именно сравнение систем машинного перевода, трудно достигнуть. В контексте информативного перевода может быть организована конкуренция между системами, работающими в одинаковых языковых парах и в тех же тематических областях. Однако это пока что сделано, как сообщает К. Буатэ, только в отношении японско-английского перевода.

В том, что касается профессионального перевода, это почти невозможно, поскольку система машинного перевода может показать свою ценность только после настройки на конкретную область (по терминологии) и стиль (по грамматике). Такая специализация весьма дорога. Даже если подобное сравнение будет организовано на справедливой основе (финансирование нескольких исследовательских групп для специализации их систем), пока нет согласия относительно критериев, которые следует при этом применять. Суждения о качестве перевода весьма субъективны. Измерения стоимости и эффективности за некоторый период времени более объективны, но требуют много времени для того, чтобы пользователь

привык к системе. При этом важным критерием является легкость ее улучшения — пополнение словарей, повышение эффективности алгоритмов и пр. С этими рассуждениями Буатэ нельзя не согласиться.

Современные системы машинного перевода

В настоящее время пользователю предоставляется множество систем машинного перевода, дающих достаточно грубый перевод, но тем не менее позволяющих преодолевать языковой барьер. П. Н. Хроменков исследовал наиболее известные из них, доступные пользователям, и показал, что с точки зрения теории практически все они так или иначе реализуют подход, который можно назвать «моделью переводных соответствий» в данной языковой паре (парах) [Хроменков 2005] (см. также [Марчук 1997] относительно модели «текст-текст» в современных исследованиях и разработках). Довольно часто современные системы машинного перевода дают некоторый полупродукт, который далее редактируется человеком-редактором или переводчиком также с использованием компьютерных средств в виде автоматических словарей или так называемой «переводческой памяти» (translation memory). В статье [Светова и др. 2002] описываются системы автоматизированного перевода со всякого рода вспомогательными средствами. Такое взаимодействие человека с информационными технологиями позволяет в первую очередь обеспечить быстрое выполнение переводов, что весьма существенно в информационную эпоху и, кроме того, дает перевод достаточно высокого, во всяком случае приемлемого для заказчика качества. В сборнике Всероссийского центра переводов «Перевод: традиции и современные технологии» дается полезный обзор существующих систем и приемов автоматизированного перевода (см. [Светова и др. 2002]).

Особо следует отметить вопросы машинного перевода с/на восточные языки. Эти языки в настоящее время выходят на мировую арену, и проблема их анализа и синтеза для машинного перевода, равно как и для других видов автоматической обработки, становится весьма актуальной. Перевод материалов этих языков интересен также и в плане сопоставительной лингвистики, поскольку различия в грамматическом строе, особенно в синтаксисе, между этими языками и европейскими значительны и требуют особых решений. В то же время, формальное изучение различий и инвентаризация способов их преодоления дают интересный материал для обобщений в рамках общей лингвистической теории и позволяют заново оценить известные лингвистические позиции и проблемы, такие, например, как части речи (см. [Кривонос 2001]). Вопросы машинного перевода с восточных языков на европейские и наоборот рассмотрены в работе Ю. В. Марчука [Марчук 2002].

МАШИННЫЙ ПЕРЕВОД КАК ЦЕНТРАЛЬНАЯ ПРОБЛЕМА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Место машинного перевода в ряду интеллектуальных задач

Системы машинного перевода текстов с одного естественного языка на другой осуществляют сложные преобразования на всех языковых уровнях двух и более языков, ставя целью передачу смысловой информации текста одного языка и создание эквивалентного ему по форме и содержанию текста на другом, выходном, языке. В данном разделе мы хотим обратить внимание на место машинного перевода в ряду других задач искусственного интеллекта. Мысль о том, что машинный перевод является центральной проблемой искусственного интеллекта, принадлежит академику Ю. В. Рождественскому, много занимавшемуся вопросами новой фактуры речи — современными способами работы с текстами в новых условиях массовой коммуникации и культуры. Можно спорить о том, действительно ли именно машинный перевод является центральной проблемой моделирования человеческого интеллекта — искусственного интеллекта, хотя многие ученые вообще скептически относятся к формуле «искусственный интеллект». Тем не менее никто не будет спорить о том, что перевод с одного естественного языка на другой действительно является сложнейшей проблемой, затрагивающей практически все аспекты человеческого мышления и способов языкового выражения мысли.

В связи с этим следует отметить новое направление прикладных лингвистических исследований — изучение естественного интеллекта. Н. К. Рябцева пишет: «...В результате естественный интеллект — *самый сложный и неисчерпаемый объект изучения* — стал, в определенном смысле, «фоновым», как бы исчерпавшим свои эпистемические возможности термином, а не «выделенным» и продуктивным понятием. Соответственно «за кадром» остается масса важнейших вопросов: почему всякое обучение, строго говоря, не так эффективно, как хотелось бы? Почему автоматизированные системы обработки информации на естественном языке, несмотря на гигантские усилия своих разработчиков, мягко говоря, далеки от совершенства? Как соотносятся такие «ментальные» явления, как сознание, подсознание, разум, рассудок, рассудительность, ум, умонастроение, познание, мышление, память, знания, взгляды, а также смекалка и т. п. и интеллект?» [Рябцева 2004]. Язык и интеллект человека неразрывно связаны и потому представляют собой важнейший источник сведений друг о друге, говорит далее автор.

Целью машинного перевода, как и всякого другого, является передача смысла. Для этого необходима экспликация смысла, в том числе и средствами языка-цели. Описание смысла сообщения в терминах смыслового языка, например искусственного, есть тоже перевод. Между тем такое описание и извлечение смысла, эквивалентное распознаванию, пониманию, есть не что иное, как центральная задача моделирования мышления. Поэтому система машинного перевода, моделирующая одну из центральных задач познания мышления, есть одновременно система искусственного интеллекта.

Представляет интерес в этой связи концепция академика Ю. В. Рождественского [*Рождественский* и др. 1988, *Рождественский* 2003]. Он выделяет элементарные, базовые системы искусственного интеллекта и сложные системы, различие между которыми состоит в назначении, характере операций и в информационном продукте. Сложные системы включают моделирующие, экспертные, обучающие, игровые системы и основаны уже не на составлении или прочтении текста по буквам, как простые системы, а на семантике текстов. Для операций над семантикой текстов необходимо сопоставление текстов и их частей, что становится возможным лишь при определенном отношении к содержанию текстов.

Переводное значение отличается от значения, выделяемого как понятийное или образно-понятийное. Переводное значение: не имеет характера названия примера, установления имени; не возникает в процессе таких операций, как филологическое истолкование текста, интерпретация стиля источника или создание стиля перевода; существует только в переводе и представляет собой особую семантическую аспектацию слова и текста, по мнению Ю. В. Рождественского.

Многие системы автоматизированного перевода присоединяют к сопоставительной переводной системе систему истолкования текста как систему моделирования содержания текста. Они содержат также словарь для справок редактора перевода. Таким образом системы автоматизированного перевода соединяются с моделирующими и экспертными системами.

Для своего функционирования системы машинного перевода нуждаются в записи, редактуре, устном вводе текста, устном управлении, анализе и синтезе текста, т. е. система автоматизированного перевода основывается на базовых системах искусственного интеллекта. Это значит, что система машинного перевода находится в центре систем искусственного интеллекта. Являясь частью информационной деятельности, она порождает тексты информатики, может быть использована для обучения, соединяется с экспертными и другими системами искусственного интеллекта. Схематически такое положение может быть изображено следующим образом (рис. 16).

Лингвисты и филологи, хорошо знающие лингвистические, текстовые трудности понимания, интерпретации, перевода, анализа и

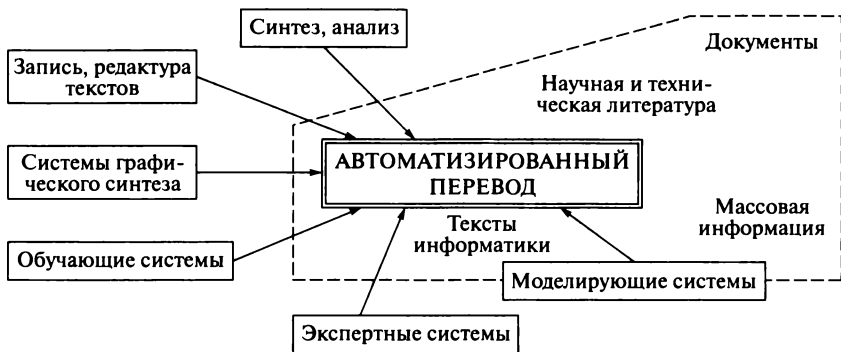


Рис. 16. Система автоматизированного перевода в ряду других систем искусственного интеллекта

синтеза естественно-языковых текстов, ставят обычно системы машинного и автоматизированного перевода в центр систем искусственного интеллекта. Дело здесь в том, что специалисты такого профиля хорошо представляют себе, какие последствия могут иметь неправильно понятое слово, не воспринятая в данной коммуникативной ситуации мысль, неверный синтаксический анализ, ошибочный синтез. Интересно отметить, что специалисты другого профиля в искусственном интеллекте считают проблему перевода каким-то побочным следствием кодирования или более широкой задачи общения с ЭВМ по тому или другому поводу. Два подхода, иллюстрирующие такую точку зрения, целесообразно рассмотреть.

Например, А. Н. Заслонко и Г. К. Хахалин [Заслонко и др. 1989] считают машинный перевод частным случаем лингвистического транслятора, рассуждая следующим образом.

Основной задачей машинного перевода является анализ и синтез текста на естественном языке, т. е. задача лингвистической трансляции. Особенностью описываемого метода трансляции является наличие модели проблемной области, к которой относятся тексты на естественном языке. Эти тексты переводятся в структуры на языке этой модели при анализе, а затем эти структуры переводятся на тексты естественного языка при синтезе. В качестве метода трансляции используется метод контекстного фрагментирования. Сам лингвистический транслятор представляет собой трехуровневую систему: лингвистический процессор, включающий анализатор и синтезатор, лингвистическую модель, включающую знания о грамматике и семантике, и ассоциированные процедуры, осуществляющие взаимосвязь декларативной и процедурной частей лингвистического транслятора. Такая структура транслятора выбрана с учетом возможностей его настройки на различные естественные языки, разные системы классификации лингвистических отношений и пр.

Лингвистическая модель разделяется на три составляющие: грамматическую, интерпретационную и модель проблемной среды. Для анализатора и синтезатора, работающих в режиме двуязычного перевода, первая и вторая составляющие различны, а третья является общей. Если лингвистическая модель для анализатора и синтезатора одна и та же, то транслятор работает в режиме перефразировки, что важно для его настройки и отладки.

Расширенный компонент анализатора предназначен для анализа сложных, осложненных, эллиптических и анафорических на естественном языке предложений. С помощью этого же метода — с использованием контекстов другого уровня — сложные предложения разбиваются на простые фразы, и строится структура сложного предложения, которая композиционно определяется взаимосвязями фраз в предложении. Каждая фраза обрабатывается базовым компонентом, а результаты перевода на язык модели компонуются с учетом связей между фразами и на основе семантических контекстов. Для восстановления эллипсисов и разрешения анафор используются дискурс и механизмы сопоставления, позволяющие найти недостающие структурные элементы для эллипсисов или антецедент для ссылки.

Каждое естественно-языковое предложение, прошедшее через анализатор, представляется структурой на языке модели. Синтезатор по этим структурам строит предложение на другом естественном языке либо с учетом структуры предложения анализируемого языка, либо без учета. В последнем случае осуществляется пересказ, т. е. перевод по смыслу, соответствующий входному тексту, но по структурному выражению не совпадающий с оригиналом. В основе синтеза также лежит метод контекстного фрагментирования.

Разработанная версия лингвистического транслятора под названием АДАЛИТ предназначена для общения на проблемно-ориентированном естественном языке конечных пользователей с интеллектуальными вычислительными системами, но может быть использована и как самостоятельная система в рамках системы машинного перевода, по мнению авторов. Проверка лингвистического транслятора осуществлялась на нескольких задачах (доступ к базам данных, кодирование медицинских диагнозов и пр.), на нескольких проблемных областях (хранение кадровой информации, медицина, геометрические объекты и пр.). Возможность использования АДАЛИТа для машинного перевода иллюстрируется проведенными экспериментами, в которых на вход анализатора поступали предложения на английском и эстонском языках, а на выходе синтезатора выдавались предложения на русском языке. Версия транслятора реализована на ЛИСП: объем — порядка 2 тысячи строк; среднее время анализа — 3 секунды на слово, синтеза — 1 секунда на слово на ЭВМ ЕС 1055 при объеме словарей порядка 500 слов естественного языка.

Расценивая это предположение, видим, что оно исходит из слишком оптимистической предпосылки о том, что словарь в 500 слов по

качеству и лингвистическим коннотациям ничем не отличается от словаря в 5000 слов, а последний — то же, что и словарь в 50000 слов. Такой взгляд кажется слишком упрощенным. Только в очень ограниченном подязыке и узкой предметной области можно уверенно работать с таким словарем. При возрастании словаря за счет употребительных, общетехнических и общенаучных слов неизбежно существенное расширение омонимии, лексической многозначности, диапазона синтаксических конструкций и разного вида неоднозначностей.

Машинный перевод как «побочное следствие» преобразований

Другая концепция ускоренного достижения машинного перевода — вывод из некоторого обобщенного представления о сущности языкового знака. А. В. Танасиенко считает, что одна общая концепция языкового знака даст возможность осуществить машинный перевод как некоторое побочное следствие общих языковых преобразований [Танасиенко 1989].

Назовем пару обобщенных объектов *a* и *b* модельной парой, если они имеют множество общих свойств и признаков, взаимно отражаемых ими. Каждый из этих объектов является моделью другого, а принцип, декларирующий относительность статуса модели, назовем принципом модельной относительности. Воспринимаемый субъектом материальный объект и его идеальный (перцептивный) образ в сознании субъекта есть не что иное, как модельная пара, т. е. пара обобщенных объектов, находящихся в модельном отношении. Автор называет *автонимным знаком* материальный элемент такой модельной пары и автонимным образом — идеальный ее элемент. В термине «автоним» объединяются два эти понятия: автонимный знак — он материален и «автонимный образ» — он идеален. Одним из простейших примеров естественно-языкового автонимного знака является фонема. Из такого определения следует, что любой доступный восприятию материальный объект можно считать автонимным знаком, если он хотя бы раз был представлен субъекту и сформировал перцептивный образ в его сознании, причем число таких представлений можно интерпретировать как некоторую характеристику его автонимной «знаковости». Материальные объекты, не удовлетворяющие этим условиям, например принципиально ненаблюдаемые или ни разу не встретившиеся субъекту, автонимными знаками не являются.

Автору этой концепции представляется принципиально возможной реализация системы, способной автоматически строить по потоку текстов адаптивную иерархию языковых автонимов, отражающую систематическую информацию, заключенную в тексте, и в этом смысле реализующую машинную процедуру понимания. Автомати-

ческое наращивание иерархической автономной структуры должно обеспечиваться за счет работы алгоритмов частотного анализа.

Вследствие концептуальной компактности естественного языка при правильном отборе алгоритмов количественный рост автономов не должен прогрессировать геометрически, но, напротив, должен зафиксироваться по достижении определенного иерархического уровня.

Система автоматического перевода с одного естественного языка на другой может быть получена в результате соединения (человеком, владеющим обоими языками) двух автономных структур, выращенных для данной пары языков автоматически. После соединения автономных структур такая система должна обеспечивать автоматический перевод для двух конкретных языков как в прямом, так и в обратном порядке. Перевод должен поддерживаться только за счет работы алгоритма, моделирующего распространение возбуждения в автономной структуре под воздействием входного текста.

За счет реализации описанного механизма система, по мнению автора данной концепции, может обойтись без таких атрибутов машинного перевода традиционного типа как жесткие алгоритмы синтаксического и морфологического анализа, а также машинные словари фиксированных словоформ, подготавливаемые человеком вручную.

Возможности такого решения

Оценивая данное предложение с точки зрения практических аспектов машинного перевода, следует отметить, что пока еще нет ни одной действующей системы машинного перевода, которая была бы построена только на основе чисто глобальных теорий, подобных этой. Дело в том, что автоматическое построение систем машинного перевода, которое заложено в подобной концепции и при котором учитываются некоторые общие свойства текстов и языковой действительности, обычно натывается на преграды, состоящие в том, что языковые структуры весьма индивидуальны и ориентированы в подавляющем большинстве случаев не на классы явлений, а на отдельные явления и языковые факты, такие, например, как отдельные слова, каждое из которых представляет собой самостоятельную проблему для анализа и синтеза и не укладывается в какие-то общие всеохватывающие схемы.

Таким образом, общезыковые преобразования не являются достаточными для того, чтобы построить машинный перевод как некоторое частное следствие из общих теорем. Как правило, проекты такого рода выдвигаются либо специалистами по логике, либо программистами, не представляющими трудности конкретной общезыковой задачи, каковой является перевод. Главная информация, подлежащая переводу, заключается не в общем, а в наборе индивидуальных языковых средств. А именно эти средства и игнорируются

общими универсальными теориями преобразований. Неслучайно важными моментами новой теоретической лингвистики провозглашены, как мы выше отмечали, основополагающие значения языковых частных и практических решений.

Системы машинного перевода АМΠΑР и СПРИНТ

Система машинного перевода, служившая прототипом системы АМПАР, создавалась коллективом разработчиков под руководством Ю. А. Моторина в течение нескольких лет, начиная с 1956 г. Эти работы начались одновременно с исследованиями по машинному переводу в Институте точной механики и вычислительной техники Академии наук СССР, где ими руководили Д. Ю. Панов (со стороны вычислительной техники и программирования) и И. К. Бельская (лингвист). Создавался алгоритм перевода с английского языка на русский, результаты работы над которым в довольно полном виде были опубликованы позднее [*Бельская* 1969]. В 1974 г., когда был образован Всесоюзный центр переводов научно-технической литературы и документации (ВЦП) Государственного Комитета СССР по науке и технике и Академии наук СССР, а в нем — отдел машинного перевода, руководство которым было поручено Ю. В. Марчуку, эта система была предоставлена в порядке технической помощи данному отделу и началось ее программирование для тогдашних серийных ЭВМ общего назначения серии ЕС (единая система), серий ЕС-1040 и ЕС-1065. Одновременно с программированием производились совершенствование словарей и алгоритмов системы и настройка ее по тематике на английские тексты по вычислительной технике и программированию. Тогда же появилось название системы АМПАР, что означает «автоматизированный машинный перевод с английского на русский». Слово «автоматизированный» как бы подразумевает, что перевод требует редактирования человеком и не является полностью автоматическим в этом смысле.

Теперь можно видеть, насколько сильно была в то время скомпрометирована идея машинного перевода чрезмерным выпячиванием концепции «передачи смысла». «Перевод без перевода, без машин, без алгоритмов» утвердился в академической науке и подавлял здравомыслие в составлении словарей, описаний и алгоритмов. Декларировалось, что составление алгоритма — не дело рук лингвистов, пусть этим занимаются математики, а лингвисты должны лишь давать описание языковых фактов. Концепции, подобные этой и языку-посреднику в виде «пучков грамматических соответствий», надолго затормозили отечественные исследования, и только с образованием ВЦП возобладали практический подход, основанный на здравом смысле, хорошем знакомстве с возможностями вычислительной техники и программирования того времени и правильной постановке цели.

В настоящее время АМПАР породила серию систем машинного перевода на персональном компьютере под названием СПРИНТ.

Глава 12

ОБЩАЯ СТРАТЕГИЯ РАЗРАБОТКИ СИСТЕМ МАШИННОГО ПЕРЕВОДА НА ОСНОВЕ МОДЕЛИ ПЕРЕВОДНЫХ СООТВЕТСТВИЙ

Общие принципы построения модели

Появление идеи машинного перевода, эвристический подход к задаче перевода с одного языка на другой породили особый взгляд на проблему получения из текста сведений, необходимых для формального осуществления этого процесса. Для передачи процесса компьютеру, для составления алгоритма требовалось прежде всего разбить перевод на отдельные последовательные этапы. Те этапы и та последовательность работы переводчика, которые описывались в моделях перевода человека, не могли быть достаточно эффективно представлены в чисто формальном виде. Трудно в виде алгоритма передать, например, проникновение переводчика в культурную среду, в которой зародился данный текст. Каковы должны быть его действия и на каком материале? Составление общей стратегии перевода, определение образа получателя перевода и другие этапы переводческих действий, описанные в «человеческих» моделях перевода, нуждались в осмыслении и более точной формализации. В достаточно окончательной, утвердившейся форме этапы формализации сложились уже в самых ранних алгоритмах машинного перевода. Довольно естественным путем выделились стадии первичной обработки (поиска входной словоформы по словарю), морфологического анализа, обработки фразеологических оборотов (фразеологизмов), анализа некоторых особенных, присущих данному конкретному языку форм, например -ing-форм в английском языке, поиска по словарю слов входного языка для нахождения переводных эквивалентов, контекстного анализа для установления типов грамматических конструкций и т. п.

Модель машинного перевода на основе переводных соответствий описана в нескольких работах [Марчук 1983, 1985, 1997]. Теоретический принцип, заложенный в эту модель, заключается в воспроизведении действий переводчика, работающего в данной языковой паре. Последовательно двигаясь от фразы к фразе, он строит в уме некоторое приближенное представление о содержании текста, затем сопоставляет это представление с языковыми средствами, выбирая переводные эквиваленты и подыскивая переводные соответствия

трех типов: эквивалентные, вариантные и трансформационные. Такая модель действий заимствована из традиционной теории перевода. Различаются также статика и динамика переводческого процесса. Статическая часть — словарь, правила грамматики, семантические закономерности. Динамическая — система определения переводных соответствий. Сами эти соответствия целесообразно находить и устанавливать на основе параллельных текстов, исходя из концепций современной корпусной лингвистики.

Как показал П. Н. Хроменков [Хроменков 2005], все современные системы машинного перевода практического направления так или иначе действуют именно на основе концепции переводных соответствий, иногда в некоторых модификациях. В частности, такими модификациями являются так называемые трансферные системы, где переводные соответствия образуют особый этап, называемый трансфером [Новиков В. А. 2001]. Совершенствование качества перевода может зависеть от введения в алгоритмы анализа и синтеза более совершенных правил, например таких, которые разработаны для системы САПФИР, однако этот путь достаточно трудоемок.

Блок-схема алгоритма машинного перевода на основе переводных соответствий

Последовательность этапов анализа и синтеза текста в системе машинного перевода, построенной по модели переводных соответствий, имеет следующий вид (рис. 17).

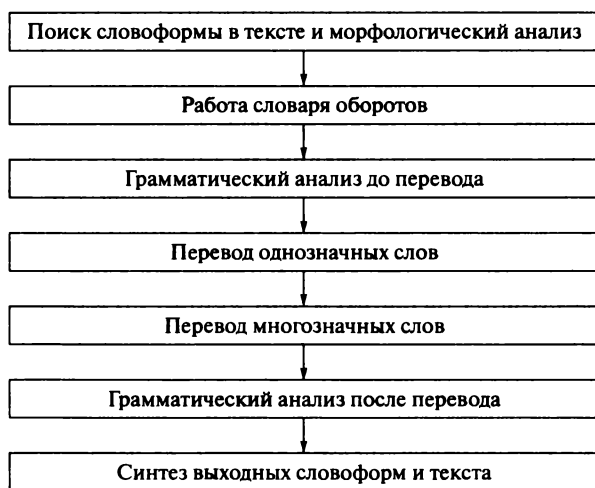


Рис. 17. Блок-схема общего алгоритма машинного перевода в модели переводных соответствий

Стратегия реализации этой модели складывается из следующих действий: подбор исходного корпуса текстов; установление переводных соответствий на разных уровнях языковой структуры; определение формальных категорий для морфологии, синтаксиса и семантики; составление алгоритмов идентификации формальных категорий; составление общего алгоритма системы машинного перевода.

В приведенной блок-схеме не указаны некоторые вспомогательные этапы, такие, как проверка правильности грамматической информации, перестановка/порядок слов и пр., которые зависят от типов сопоставляемых языков.

Данная схема была применена к разработке систем АМПАР, НЕР-ПА (немецко-русский автоматический перевод), СПРИНТ (Всесоюзный центр переводов). Кроме того, по этой схеме разрабатываются системы перевода для таких языковых пар, как англо-персидский и русско-персидский [Али Реза Валипур 1998, Мохаммад Реза 1998, Эмарлу Рамезанли 1998, Магомедова 2004, Тайеби 2002], а также для перевода с русского языка на китайский [Чжань Дэхуа 1994].

Данная общая схема, как показывает опыт, достаточно хорошо подходит к языкам различного строя. При этом следует учесть, что реализация ее должна проходить по общим законам корпусной лингвистики. Корпусная лингвистика, базирующаяся на больших массивах языковых текстов, позволяет извлекать объективные лингвистические данные и делать содержательные выводы на их основе с высокой степенью достоверности, что гарантирует эффективность работы системы машинного перевода, построенной с учетом этих данных. При этом важен учет параллельных текстов, т.е. оригинала и перевода.

Последнее особенно важно. Первые работы с параллельными текстами были выполнены в конце 80-х — начале 90-х гг. прошлого столетия в рамках создания различных систем статистического машинного перевода [Рахимбердиев 2002, Gaussier et al. 2000]. Была создана первая система, извлекающая знания о языке оригинала, языке перевода и правилах перевода исключительно из массивов примеров перевода. Эти работы вызвали критику со стороны «традиционной лингвистики», так как в них отвергались подходы к машинному переводу, считавшиеся общепринятыми в течение десятков лет. Несмотря на недостатки этих систем, статистический подход к машинному переводу привлек к себе внимание колоссальным сокращением трудоемкости их построения по сравнению с системами традиционными. Их неоспоримым достоинством является отказ от ручного составления машинных словарей и грамматик; если в логике системы обнаруживается ошибка, ее устранение в худшем случае означает необходимость повторного запуска процедуры извлечения параметров из корпуса примеров, а не ручное переписывание этих ресурсов.

Первые методики машинного перевода на основе примеров послужили толчком для дальнейших исследований в области извлечения лингвистической информации из параллельных текстов. С одной стороны, развивались сами системы статистического машинного перевода за счет разработки более сложных лингвистических моделей, отказа от идеи абсолютной независимости алгоритмов от обрабатываемых языков, привлечения морфологического анализа и т. п. Так, хотя в описанных системах статистического перевода не используются развитые формальные грамматики, понятие класса слов уже стало частью таких алгоритмов (см. также [Кривоносов 2001]). С другой стороны, произошла некоторая переоценка роли параллельных текстов. Алгоритмы обучения на параллельных корпусах стали применяться в инструментах, облегчающих труд переводчика, а также в системах автоматической проверки переводов, выполненных человеком [Марчук 2002].

Таким образом, новая ветвь прикладной лингвистической науки, а именно корпусная лингвистика, которая имеет свою методику, цели и задачи обработки естественно-языковых текстов, вносит существенный вклад в создание эффективных современных систем машинного перевода.

Различие в подходах к моделированию перевода

Приведенная выше модель перевода на уровне переводных соответствий представляет собой один из подходов к моделированию перевода. Появление идеи машинного перевода, эвристический подход к задаче перевода породили особый взгляд на проблему получения из текста сведений, необходимых для формального осуществления этого процесса. Обратим внимание на то, что многочисленные модели «человеческого» перевода, рассмотренные нами выше, плохо поддаются формализации. Так, например, этапы действий переводчика определены практически во всех приведенных выше моделях, однако если поставить вопрос о том, какие формальные границы разделяют эти этапы, то ответ получить будет довольно трудно. Например, начальным этапом является определение стратегии перевода. Как описать этот этап в виде алгоритма, состоящего из последовательности действий, каждое из которых осуществляется на определенном материале, дает конкретный результат, отграничено от других действий? Как точно определить менталитет (возможного) потребителя? Какими параметрами можно точно и непротиворечиво оценить адекватность перевода с учетом всех возможных условий его функционирования? Список таких вопросов можно было бы продолжить.

Для передачи процесса компьютеру и составления алгоритма требуется прежде всего разбить перевод на отдельные последова-

тельные этапы. Впрочем, в достаточно окончательной, утвердившейся форме они сложились уже в самых ранних алгоритмах машинного перевода. Довольно естественным путем выделились стадии первичной обработки (поиска по словарю входных словоформ переводимого текста), морфологического анализа, обработки фразеологических оборотов, анализа особых форм переводимого языка (например, анализа -ing-форм английского языка), нахождения переводных эквивалентов однозначных слов, запросы контекста для перевода многозначных слов, определение грамматической информации и т. п. (см. блок-схему алгоритма системы АМПАР). Следует заметить, что между этими этапами и действием человека-переводчика нет прямой аналогии. Можно утверждать, таким образом, что в начале моделирования перевода для целей перевода машинного было принято довольно четкое и однозначное деление процесса перевода на анализ, синтез и некоторый промежуточный этап, который в разных системах назывался по-разному и имел довольно различные функции в машинной обработке, однако в целом служил одной цели, а именно поиску переводных соответствий. В этот этап можно включить поиск по словарю, разрешение неоднозначности слов с помощью анализа контекста, анализ отдельных видов неоднозначности, поиск необходимых соответствий синтаксических структур и т. п. Впоследствии этот этап получил название «трансфер».

Моделирование перевода в традиционном языкознании и переводоведении и для машинного перевода шло разными путями. Машинное моделирование в целях машинного перевода не отличало перевод от языковой деятельности другого рода, например от автоматического извлечения содержания текста на естественном языке. Перевод моделировался, как любой языковый процесс, с теми же стадиями и с теми же единицами анализа, как и обычный разбор предложения. Такое моделирование шло по пути «анализ-синтез» или «анализ-трансфер-синтез». При этом собственно переводные соответствия занимали места на периферии перевода. В отличие от этого в моделировании перевода как «человеческого» процесса филология и лингвистическая наука ставили на первое место моделирование процесса установления переводных соответствий. Даже чисто поверхностный анализ многих работ по теории перевода, имеющих большое значение и вносящих существенный вклад в науку о переводе, показывает, что после вводной теоретической схемы обычно следуют обширные и весьма полезные данные и наблюдения о способах перевода отдельных достаточно общих и распространенных, но также и достаточно статистически редких явлений. Здесь можно отметить некоторые парадоксы человеческого и объективного статистического восприятия. Так, интуитивно представляется, что в английском языке глаголы в пассиве распространены по текстам достаточно часто; статистический же подсчет на большом материале

показывает, что всего 8 % английских глаголов в текстах употреблены в пассиве.

Анализ естественного языка с целью построения алгоритмов его обработки естественно вызвал к жизни вопрос о единицах анализа. Перевод поставил вопрос о том, существует ли единица перевода.

Единица перевода

Можно начать с мнений основоположников теории перевода, которые (мнения) были высказаны достаточно давно, но продолжают оставаться актуальными в теории «человеческого» перевода. «Думается, что само понятие «единица перевода» представляет собой противоречие в терминах прежде всего потому, что любая единица — это постоянная величина, тогда как так называемая «единица перевода» по определению является величиной переменной. Кроме того, в лингвистике любая единица представляет собой постоянную величину, образующую тот или иной уровень языка. В то же время в теории перевода за единицу принимаются варьирующие и неопределяемые в лингвистических терминах речевые отрезки исходного языка. Наконец, любая единица позволяет как бы «измерить» однородные величины, представив их в виде линейной последовательности или совокупности определенных единиц. Например, высказывание может быть представлено как последовательность фонем или морфем. Что же касается процесса перевода, то он никак не может быть представлен как простое соединение единиц. Здесь имеют место куда более сложные операции» [Швейцер 1973]. Я. И. Рецкер более определенно отзывается о единице перевода, утверждая, что ее вообще не существует. «Несомненно, выделение определенных «единиц перевода» имеет большое значение для машинного перевода, а также для синхронного и последовательного устного перевода... единица перевода зависит от той пары языков, которые участвуют в переводе: то, что является единицей при переводе с русского на французский, может не быть единицей при переводе на немецкий... Фактически в процессе письменного перевода единицей перевода может быть и слово, и словосочетание, и синтагма, и целое предложение, и абзац, и весь переводимый текст... сопоставление различных точек зрения убеждает в том, что для письменного перевода такой проблемы (т. е. единицы перевода. — Ю. М.) не существует» [Рецкер 1974]. Л. С. Бархударов отмечает, что на каждом уровне есть свои единицы перевода. «Под единицей перевода мы имеем в виду такую единицу в исходном тексте, которой может быть подыскано соответствие в тексте перевода, но составные части которой по отдельности не имеют соответствий в тексте перевода. Иначе говоря, единица перевода — это н а и м е н ь ш а я (минимальная) языковая единица в тексте на исходном языке, которая имеет соответствие в тексте на переводящем языке; она, как

мы увидим дальше, может обладать сложным строением, то есть состоять из еще меньших единиц исходного языка, но части ее, по отдельности взятые, «непереводимы», то есть в тексте перевода им никаких соответствий установить нельзя, даже если в исходном языке они обладают своим собственным, относительно самостоятельным значением». И далее: «По сути дела единицей перевода может быть единица любого языкового уровня. Поэтому необходимо, прежде всего, выяснить, какие уровни языковых единиц выделяются в структуре языка вообще» [Бархударов 1975].

Следует подчеркнуть, что для переводчика важно не только то, что в структуре текста выделяются какие-то единицы, а главным образом то, каким образом эти единицы связаны с переводом. Рассмотрим следующую точку зрения.

«Чтобы переводчик правильно понял смысл переводимого отрезка текста, единица коммуникации (или интерпретации) должна быть функциональной, так как он не занимается лексико-грамматическим сопоставительным анализом. Верно, что основа такой единицы должна быть лингвистической, однако необходимо также принимать во внимание «текстуальные функции», такие, как «концептуальные отношения», телеологию и структуру сообщения и другие металингвистические факторы. Переводчик не интересуется чисто механическим разделением текста или обнаружением посредством какого-либо вида тагмемного анализа единиц формы и содержания для статистических целей. Он заинтересован в функциональной единице, а не в грамматических средствах, которые ее образуют. Подчинение значения синтаксическим структурам приведет к искаженной интерпретации».

«Следует начинать от свидетельства того, что существует одноднзначное соответствие между мыслью и словом или между образом мысли (thought pattern) и синтаксическим образцом (syntactic pattern) и что не существует определений не-синтаксических единиц дискурса или предложения» [Vazquez-Ayora 1982].

Такие соображения представляются важными. Действительно, в поисках эквивалентного языкового выражения данного содержания на языке оригинала переводчик не имеет права уходить слишком далеко от дискурса — от текста в его полном объеме. Всякие возможности перефразирования на самом деле в переводе чрезвычайно ограничены. Передача содержания своими словами, т.е. интерпретация, в общем случай частный, вынужденный, и брать его за основу переводческого процесса можно только тогда, когда соответствие на уровне языкового выражения отсутствует. Между тем некоторые семантические теории перевода, базирующиеся на «глубинных структурах», делают такой прием чуть ли не основой всякого перевода. Это безусловно неверно.

До сих пор продолжают поиски единицы перевода как некоторого исходного понятия, которое могло бы лечь в основу теории пе-

ревода или теории ряда последовательных преобразований, направленных к восприятию содержания текста. В советских работах (Л. С. Бархударов, А. Д. Швейцер, В. Н. Комиссаров и др.) даются обзоры известных представлений о единице перевода. Иногда она может совпадать с единицей анализа и/или синтеза, например с морфемой, однако круг таких единиц довольно ограничен. Достаточно часто единицей перевода может выступать слово, однако оно не исчерпывает всех возможностей и всех требований, предъявляемых к единице перевода со стороны «человеческой» теории перевода. Наибольший интерес теории перевода вызывают единицы, так или иначе связанные с восприятием частей текста, значительно больших, чем слово.

Практически переводчик постоянно имеет дело с единицей перевода, однако как теоретическую сущность определить ее чрезвычайно трудно. «Одна из особенностей перевода как вида коммуникативной деятельности состоит именно в том, что он нацелен не на абстрактные сопоставления языковых единиц соответствующих языков, а на адекватное воссоздание содержания подлинника. Словарь, например, является наихудшим помощником переводчику в том плане, что он, как правило, никогда не дает необходимых контекстуальных соответствий, а в лучшем случае перечисляет некоторые номинативные значения слов и выражений».

«Контекстуальная зависимость значений языковых единиц как отдельных предложений, так и всего текста представляет собой то необъятное поле творческой деятельности переводчика, которое дает основание относить процесс перевода скорее к искусству, нежели к науке» [Колшанский 1980]. Теоретики перевода постоянно подчеркивают связь перевода с четким представлением о смысле переводимого текста в целом (см. выше обзор существующих моделей «человеческого» перевода). Чтобы перевести, нужно понять, после чего единица перевода выделяется как бы сама собой. Чрезвычайно важный аспект поиска этой единицы зависит также и от некоторых общих положений, которые, может быть, можно было бы отнести к стилю или другим общим характеристикам высказывания. Буквализм, сохранение черт языка оригинала в переводе, воспринимается носителями языка как серьезные отклонения от норм языка выходного. На самом деле следует учесть, что имеет место сближение культур. Самый точный, по мнению Ю. В. Рождественского, есть перевод буквальный. Однако до тех пор, пока он будет признан именно переводом, а не набором ошибок с точки зрения выходного языка, пройдет достаточно большой период приработки и выравнивания языковых культур. Поясим это примером из области новых информационных технологий. Слова *инфраструктура*, *файл*, *терминал*, *процессор* и пр. в первых вариантах перевода, когда переводчику и специалисту, не имевшему возможности поставить в соответствие термину оригинала термин перевода или хотя бы эквивалентное понятие по

причине полного отсутствия денотата, приходилось давать кальку или описательный вариант, воспринимались читателями перевода как грубые ошибки переводчика. Лишь по прошествии достаточного количества времени, когда соответствующие устройства были созданы, эти слова вошли в русский язык и стали восприниматься как не противоречащие языковым нормам. Видимо, в диахронии такие же или сходные процессы имеют место и в грамматических структурах, оборотах речи и т. п. Является ли единицей перевода предложение? Для перевода необходимо понимание, которое не может быть полным, если предложение рассматривается вне контекста. Как видно из приведенных ранее определений В. А. Звегинцева по поводу одинакового смысла предложений относительно охотника, убившего волка, отдельно взятое предложение имеет не смысл, а псевдосмысл. Любое предложение имеет неполный смысл, утверждают другие специалисты по переводу, поскольку его полный смысл невозможно вычислить, не обращаясь к предыдущему или последующему предложению или другим частям текста. Можно отметить, что предложение является и одновременно не является единицей перевода. Оно может быть единицей перевода, если понят смысл, разложено на отдельные части и каждой из них может быть поставлено в соответствие некоторое законченное высказывание в языке перевода. Оно не может быть единицей перевода, если для его понимания требуется обращение к предыдущей или последующей части текста, к соседним предложениям. Другими словами, для синтеза при переводе предложения достаточно, а для анализа — как и для перевода — его недостаточно, должна использоваться гораздо более крупная единица, связанная с разбором текста в целом.

Основной вид «понимания» в системах автоматического анализа текстов — понимание «сверху». Это есть нахождение в тексте или в одном из его промежуточных представлений тех единиц и связей, которые заложены в знаниях. Нахождение может осуществляться в двух режимах: жестком и полужестком. В жестком режиме нужная единица задается своим лексическим видом и в таком же виде (допускаются грамматические вариации) должна быть найдена в тексте. Результат представляет собой буквально «вырезку» из текста. Пример жесткого задания — термины информационно-поискового тезауруса в информационно-поисковых системах [Шемакин 1972]. В полужестком режиме единица определяется некоторой «открытой» формулой, часть элементов которой задана лексически, а на пустые места, «слоты», должны встать какие-то элементы текста, удовлетворяющие перечисленным в слотах условиям. Пример полужесткого задания — фреймы, трафареты, прототипы, сценарии, шаблоны и т. п.

При жестких и полужестких знаниях остается неузнанной часть информации, которая не попала в слоты, фреймы и т. п. Между тем эта информация и является новой, для которой и создается текст.

Слово как единица перевода

Является ли слово единицей перевода? Слово в принципе многозначно, но при переводе его происходят следующие явления. Поскольку два языка, даже тесно связанные родством, никогда не показывают даже возможности одно-однозначного перевода соответствующей лексики, чрезвычайно часто перед переводчиком стоит задача перевода многозначной лексической единицы языка-источника на выходной язык в отсутствие точно соответствующего лексического эквивалента. Но, хотя лексическая полисемия по определению включает одну лексическую единицу с более чем одним значением, большинство текущих примеров на полисемию связано только с одним значением в каком-либо данном контексте, поэтому отсутствие соответствующей полисемантической единицы в выходном словаре обычно не имеет большого значения. Многозначность единицы перевода — слова, — таким образом, разрешается в большинстве случаев контекстом. Это, безусловно, не исключает положений, в которых никакой контекст не может разрешить омонимию (многозначность). Однако в моделировании перевода важно, какова распространенность явления.

Так или иначе в языке и особенно в речи проявляются закономерности, которые могут служить определяющими в удовлетворительном моделировании перевода. Ни один из переводов не обходится без сознательных или несознательных ошибок, т.е. текст на выходном языке содержит так или иначе ряд отступлений от оригинала. При межязыковом преобразовании (как и при всяком другом виде преобразований) неизбежны потери, то есть имеет место неполная передача значений, выражаемых текстом подлинника. Стало быть, текст перевода никогда не может быть полным и абсолютным эквивалентом текста подлинника; задача переводчика заключается в том, чтобы сделать эту эквивалентность как можно более полной, то есть добиваться сведения потерь до минимума, но требовать «стопроцентного» совпадения значений, выражаемых в тексте подлинника и тексте перевода, было бы абсолютно нереальным [Бархударов 1975]. Здесь следовало бы отметить, что поскольку мы постулируем всегда, что нет языков, в которых нельзя было бы выразить какие-то значения или их оттенки, то в идеале нет двух текстов на разных языках, которые не могли бы полностью соответствовать друг другу. Именно различие культурных фонов и трудности поиска эквивалентных средств языкового выражения приводят к неполноте соответствий подлинника и оригинала. С этой точки зрения следовало бы, наверное, говорить о том, что стопроцентное совпадение значений перевода и оригинала в принципе возможно, но это некоторое идеальное состояние.

В науке подобного рода абстракции известны: это несжимаемая жидкость, абсолютно твердое тело, идеальный газ и т.п. Реальные

объекты могут с достаточной степенью точности соответствовать идеальным. Таким образом, поскольку потери информации в любом переводе реально неизбежны, нужно учесть возможность ошибиться в некоторых пределах. Вследствие этого, как представляется, к переводу применимы количественные меры, которые получили распространение в автоматической обработке текстов. Количественные измерения наиболее показательны и объективны и имеют качественное значение именно в отношении к лексическому уровню, уровню слов. Это означает, что приведенное выше соображение о том, что большинство многозначных слов разрешает свою многозначность в контекстах, может получить количественную меру. Практические исследования показывают, что количество многозначных слов, которые не разрешают свою многозначность в контексте, сравнительно невелико. Это значит, что на основе достаточно представительного конкорданса, т.е. словаря, в котором слова даны в контекстах их употребления в рамках исходного массива текстов, чем успешно занимается корпусная лингвистика, можно отделить многозначные слова от однозначных и реально перечислить все случаи, когда контекст (четко определенный) дает возможность разрешить многозначность. Здесь возникает возможность использовать активно разрабатываемый в настоящее время арсенал способов корпусной лингвистики [Мартыненко 2006].

Слово в качестве исходной единицы перевода имеет то преимущество, что возникает возможность применить количественные методы. Ни для одной другой единицы — словосочетания, предложения, синтаксической конструкции — нет такой четко определенной возможности. Безусловно, можно считать эти языковые и речевые сущности на определенных массивах текстов, но они вряд ли могут конкурировать с таким четко определенным множеством, как множество слов. Возможно применять количественные меры к таким языковым единицам, как морфемы и грамматические формы слов, однако в качестве единиц перевода такие сущности выступают крайне редко. Таким образом, мы можем ориентироваться на слово как на одно из оснований для выбора единицы перевода.

В использовании слова как единицы перевода есть ряд преимуществ. В большинстве случаев его многозначность разрешается контекстом. Будучи подверженной количественным измерениям, распространенность слова может определять границы подязыков, моделирование перевода для которых с точки зрения автоматической обработки наиболее перспективно. В рамках отдельных подязыков можно четко определить типологию перевода. Наконец, слово четко связано с таким способом организации лингвистической информации, как словарь. Он может быть построен любым способом, в том числе и с приведением сколь угодно сложной информации других уровней. Например, к словам можно присоединять отдельные синтаксические конструкции, семантические фреймы и другие шаблоны

или более сложные единицы анализа, синтеза или перевода. В то же время, на слово можно наиболее эффективно ориентировать переводное соответствие, оно легче всего механически обнаруживается в тексте, что дает возможность применять методы автоматического анализа и синтеза. Наконец, слово является носителем морфологической информации, которая, в свою очередь, может существенно помочь в раскрытии грамматической, синтаксической и семантической структуры предложения в целом. Вокруг него в тексте целесообразно организовать поиск переводного соответствия как малого, так и более крупного масштаба. Во всяком случае, если мы отказываемся от слова как единицы переводного соответствия, то это никак не освободит нас от необходимости неоднократно и с разными целями обращаться именно к нему, если мы осуществляем перевод, а не интерпретацию, смысловой анализ и т. п. Но и в этом последнем случае без обращения к слову как главному носителю информации не обойтись. Точнее, если мы выберем единицу перевода или опорный элемент переводного соответствия крупнее, чем слово, то все равно придется обращаться к нему, даже в качестве составляющего более крупного опорного элемента.

Таким образом, есть основания считать слово необходимым опорным элементом единицы перевода, которая может быть, согласно мнению теоретиков перевода, различной по величине. Таким образом, завершая рассуждения о единице перевода в теории перевода и в моделировании перевода, приходим к следующим выводам.

1. Единица перевода существует. Ее существование необходимо для моделирования перевода в прикладных целях.

2. Единица перевода есть многоуровневое образование. Она не соответствует единицам анализа ни на каком из уровней системы языка, если мы рассматриваем вопрос об автоматическом лингвистическом анализе. Это особая единица, связанная с особым лингвистическим статусом перевода, который не является смысловым реферированием, а обязан обеспечить передачу содержания при эквивалентности средств языкового выражения.

3. Единица перевода есть элемент статики переводческого процесса. В динамике ей соответствует переводное соответствие. Она может быть задана в словаре; переводное соответствие в словаре задано быть не может, оно создается в конкретном тексте.

4. Опорным элементом единицы перевода может и должно быть слово.

Еще о статике и динамике переводческого процесса

Целью перевода является передача содержания на уровне языковых средств. Заглубление в смысл (содержание) необходимо, но только до такой степени, чтобы понять, чем вызвано использование данных языковых средств в оригинале и какие языковые средства долж-

ны быть использованы в переводе. Толкование и перевод текста на выходной язык не должны переходить эти границы. Понимание в конце концов предполагает, что реципиент должен предпринять какие-то действия для восприятия текста. В таком смысле понимание выходит за рамки того, что требуется для перевода. Это существование двух текстов, коммуникативно эквивалентных: неоднозначности оригинала (с точки зрения вышеприведенного значения понимания) должны быть переданы и в переводе, а не быть «раскрытыми» и «растолкованными».

Теоретически перевод не может быть лучше оригинала. В противном случае это не перевод. «Семантический аппарат помогает также и в разрешении неоднозначностей. Однако проект машинного перевода должен также быть готов и к использованию других средств. Роль экстралингвистических данных не должна переоцениваться; иногда понимание естественного языка объявляется как *conditio sine qua non* в машинном переводе. Это можно считать преувеличением. Понимание в конечном счете означает: быть в состоянии действовать в соответствии с тем, что было сказано, действовать вне сферы языка, так сказать (под таким действием, кроме всего, можно понимать ответы на вопрос). Перевод, однако, не заходит так далеко — в собственном смысле он должен воспроизвести устно или письменно только то, что сказано или написано на входном языке. Поэтому сохранение неоднозначности входного текста в выходном должно рассматриваться скорее как достоинство, чем недостаток перевода. Читатель, обычно хорошо разбирающийся в картине мира, о которой идет речь, может прекрасно разрешить вопрос, иногда даже не подозревая, что имела место неоднозначность. Конечно, существуют такие типы неоднозначностей, которые невозможно разрешить таким путем, однако обычно такие неоднозначности нелегко разрешить каким-либо другим путем вообще» [*Kirschner* 1982].

Языковые средства можно представлять себе как некоторое упорядоченное собрание, инвентарь слов, морфем, словосочетаний, правил образования выражений языка и их преобразований, правил истолкования. Составляющие такого собрания могут быть перечислены, перенумерованы, записаны в каком-то порядке в хранилище данных. Вместе с тем, при распознавании языковых средств выражения в текстах мы должны обращаться в хранилища подобного рода за привлечением конкретных, в данный момент нужных нам сведений или данных — списков, отдельных правил анализа и/или синтеза, синтаксических конструкций и т. п. Поэтому можно сказать, что в моделировании переводческого процесса целесообразно разделять статику и динамику.

Такое разделение позволяет, во-первых, применить к упорядочению лингвистических сведений методы, разработанные и используемые в работе с базами данных в современных новых информационных технологиях, а во-вторых, создавать специальные алгоритмы

работы с этими данными. При этом следует обращать внимание на лингвистическую сторону таких алгоритмов, причем здесь возможно использовать методы, разработанные в эвристическом моделировании мышления.

В рамках такого подхода мы можем определить отношения между единицей перевода и переводным соответствием. Единицу перевода мы ранее определили как основу переводного соответствия, понятия заведомо текстового. Отсюда следует, что аппарат его нахождения должен быть помещен в динамику процесса. Единицу перевода, базирующуюся на словах или различных комбинациях слов и других лингвистических признаков, также связанных со словами, целесообразно рассматривать как часть системы статики перевода, которая может содержаться в базах данных.

Многие проблемы переводчик не решает методом простого перебора возможных решений. Поэтому само по себе разделение данных и алгоритма работы с ними еще не обеспечивает оптимального использования данных. Дело здесь не только в алгоритме, который должен быть оптимальным с точки зрения обращения к базам данных. Сами базы данных должны быть построены таким образом, чтобы можно было оптимизировать алгоритм путем упорядочения данных по отношению к точному типу алгоритма и его построения, точно отражающего данный процесс по отношению к структуре данных. Точнее, динамика определяется статикой, а статика — динамикой. Это соотношение и составляет задачу построения эффективной модели перевода, на основе которой можно создать эффективную модель перевода машинного.

Человеку-переводчику и редактору удается осуществить достаточно быстрый и эффективный поиск в огромных массивах данных, которыми он располагает, начиная перевод или редактирование. Его действия точны и целенаправленны, путь к лучшему варианту он ищет если и посредством перебора, то весьма целенаправленного. Можно предположить, что человек использует чрезвычайно эффективную и меняющуюся в зависимости от различных факторов шкалу признаков.

Концепция модели переводных соответствий для машинного перевода

Для того чтобы можно было в рамках конкретных языковых пар оптимальным образом строить переводные соответствия и их типологию, чрезвычайно важны сопоставительные исследования, в том числе и те, которые проводятся в методике традиционного сопоставительного языкознания. На этом пути можно переходить от простых соответствий ко все более сложным. Несмотря на то что теория языковых универсалий может мало что дать машинному переводу и при-

кладным задачам обработки текста, ее исходные положения могут быть частично использованы для типологических исследований в рамках отдельных языковых пар именно с целью нахождения переводных соответствий.

Методика нахождения в тексте переводных соответствий должна быть специально разработана. Создается порядок в наборе шаблонов, которым пользуется переводчик, производя последовательный перебор вариантов перевода частей предложения или предложений в целом. Ясно, что этот набор не может априори быть исчерпывающим и полным, в противном случае задачи перевода вообще не существовало бы. Он должен постоянно пополняться и не только в смысле соответствующего банка статических данных, но и в смысле алгоритмов, методик и принципиальных подходов к решению новых текстовых проблем. Система должна быть открытой во всех существенных для перевода частях.

В соответствии со сказанным выше экспликация семантики служит главным образом и в первую очередь задачам разрешения неоднозначностей предыдущих уровней анализа. Эта семантика, естественно, может и должна быть как лингвистической, так и экстралингвистической. Однако при этом очевиден приоритет семантики лингвистической.

Одним из главных способов разрешения неоднозначности остается контекст. Способы извлечения из него максимальной информации, которыми так легко и неосознанно (в ряде случаев) пользуются переводчик и редактор, должны быть эксплицированы для осуществления перевода на ЭВМ. Эта часть работы, видимо, является наиболее трудной. Попытки формализовать контекст в терминах крупных классов лингвистических явлений оказываются малорезультативными, в связи с тем что слишком много переводных соответствий связано со словом и отдельными лексическими единицами. Поэтому формализация контекста должна по необходимости осуществляться перечислением отдельных языковых фактов разной степени общности. Примером такого рода формализации является контекстологический словарь для перевода многозначных слов.

Грамматика для систем машинного перевода должна иметь специальный вид. В настоящее время в большинстве случаев форма представления основных правил совпадает с грамматикой непосредственно составляющих. Порядок развертывания трансляционных правил чрезвычайно важен для грамматической части лингвистического обеспечения машинного перевода.

Статистические данные используются для упорядочения системы нахождения переводных соответствий. Особое значение имеют данные лексического уровня. Имплицитно присутствующая статистика подъязыков, разных явлений лингвистического характера, образующих их состав, имеет безусловно определяющее значение для составления словарей и алгоритмов анализа и синтеза в рамках подязы-

ков. Общеязыковая статистика, особенно лексики, также важна для упорядочения правил и алгоритмов. Несмотря на то что статистика более сложных уровней трудно доступна и малопоказательна, например статистика синтаксических конструкций, она тем не менее может служить удовлетворительным ориентиром в разработке грамматики и анализирующей части алгоритма перевода.

Немаловажным практическим фактором в создании и эксплуатации современных систем машинного перевода является вопрос участия человека в совместной работе с ЭВМ. Здесь необходимо учесть два аспекта. Во-первых, человек при эксплуатации системы должен вносить изменения и пополнения в словари и алгоритмы анализа и синтеза. Это означает упомянутую выше открытость системы, понятность ее основных положений и доступность файлов системы для ревизии и пополнения. Очевидно, что пополнять и исправлять придется не только словари, что относительно просто, но и алгоритмы анализа, что значительно труднее. Если бы заранее знать, какого рода трудности система взять не в состоянии, можно было бы более эффективно спланировать весь алгоритм. Однако трудно априори разграничить простое и сложное для системы. Дело в том, что она с одинаковой готовностью берется решать как простые, так и сложные задачи.

Второй аспект заключается в организации оптимального взаимодействия ЭВМ с человеком в режиме промышленного ее использования для перевода, т. е. в режиме интер-, пост- или предредактирования или же во всех этих режимах сразу. Никто в настоящее время не представляет себе практического использования компьютера для перевода без редактирования.

Перемещение компонентов модели при перемене режима также очевидно: в трансляции первым работает динамический компонент, в генерации основное внимание уделено созданию предметного. Тесное взаимодействие обоих компонентов отражено взаимным направлением стрелок. Следует также подчеркнуть момент взаимодействия человека с ЭВМ: хотя на схеме от человека идут стрелки управления, такое использование может быть только весьма ограниченным (рис. 18). Поэтому удобство переводческой работы тандема «человек — ЭВМ» должно обеспечиваться конструкцией лингвистического, математического и программного обеспечения.

В такой ситуации неперенным условием практической выгоды машинного перевода является достаточно высокое качество



Рис. 18. Генерация текста в модели переводных соответствий

собственно перевода, которое снизило бы стоимость постредактирования, поскольку в большинстве случаев редактор работает в машинном переводе именно в таком режиме. Качество перевода, в свою очередь, в большой степени зависит от полноты словаря и правильности перевода лексики, как это было неоднократно показано в исследованиях, посвященных качеству машинного перевода. Однако вторым по важности, а может быть, и первым с точки зрения полноты рабочей гипотезы, заложенной в теоретическую модель машинного перевода, является принцип синтаксического анализа, который (анализ) отвечает за большую часть разрешения неоднозначностей, препятствующих нахождению переводных соответствий.

Опыт работы систем машинного перевода в ВЦП — АМПАР, НЕРПА и ФРАП — позволяет подвести некоторые итоги эффективности синтаксического анализа в системах, ориентированных на нахождение переводных соответствий.

Модель машинного перевода на основе переводных соответствий состоит из двух компонентов: предметного и динамического. Предметный компонент отвечает на вопрос о том, что должно быть переведено, а динамический — как. Модель перевода на основе переводных соответствий можно представить себе в виде некоторого автомата, имеющего два основных состояния: генерацию и трансляцию. Состояние генерации определяется двумя причинами. Во-первых, модель создается изначально на базе некоторого исходного набора языковых данных и переводных соответствий для данной пары языков. Во-вторых, при установке на переводные соответствия для каждого подязыка внутри данной языковой пары необходима точная настройка системы, ее словарей и алгоритмов анализа и синтеза на специфику данного подязыка. Этап генерации, таким образом, имеет две основные задачи: исходное состояние системы — основной набор языковых данных, выявленных при нахождении переводных соответствий в исходном конкордансе и параллельных текстах, и более точная настройка — второй этап — на лексику, синтаксис, семантику данного конкретного подязыка. Этап генерации по своему содержанию характеризуется преобладанием систематизации элементов описания: слов, единиц перевода, синтаксических конструкций, закономерностей преобразования. Здесь создается предметное наполнение системы языковым материалом.

Этап трансляции есть собственно работа системы по переводу текста с входного языка на выходной. В нем также присутствуют элементы пополнения словарей и алгоритмов системы, однако они вторичны по отношению к основной операции этого этапа — переводу.

В обоих этапах непременно участвует человек. В первом случае он действует в качестве составителя системы, хотя ЭВМ также может принимать в этом существенное участие. Во втором случае он выступает как пред-, интер- или постредактор, корректор и пр. Его участие может быть также и минимальным, если машинный перевод

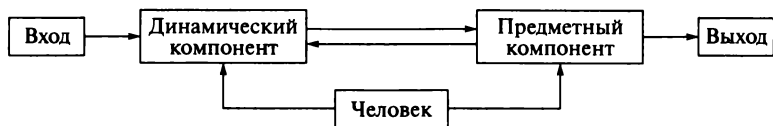


Рис. 19. Работа модели переводных соответствий в режиме трансляции

используется без редактирования как сигнальная информация. В режиме трансляции человек также выполняет функции составителя при переходе на другой подъязык, пополнении словарей и алгоритмов и пр.

Общую картину модели можно представить следующим образом (рис. 19, 20).

Несколько наиболее известных систем машинного перевода рассмотрены нами ранее. В настоящее время создаются новые системы, расширяется сфера действия уже существующих. Можно назвать следующие общие направления, по которым идет развитие действующих систем машинного перевода и соответствующих прикладных исследований.

Специфика перевода, его отличие от смысловых преобразований входного текста приводят в большинстве случаев к тому, что основное внимание уделяется поиску способов наиболее четкого и приоритетного выявления переводных соответствий на поверхностном уровне. Смысловой глубинный уровень все чаще рассматривается как источник признаков, которые используются лишь тогда, когда нет возможности разрешить неоднозначность на поверхностно лежащих уровнях анализа.

В центре моделирования ставится понятие единицы перевода, как правило, наиболее тесно связанной со словом или соответствующей лексической единицей (оборотом, словосочетанием и пр.). На этом строится типология переводных соответствий.

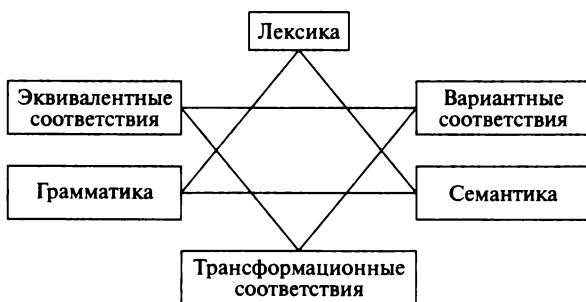


Рис. 20. Структура предметного компонента в модели переводных соответствий

Человек подает управляющие сигналы на оба компонента; на самом деле имеет место также и обратная связь, в ходе которой он проверяет результаты совершенствования и работы обоих компонентов. ЭВМ также участвует в составлении предметного компонента и в корректировке обоих компонентов по результатам работы. Это участие сводится не только к составлению конкордансов и вспомогательным работам. Для ЭВМ вполне возможно вносить существенные изменения и пополнения в состав алгоритмов, словарей, математического и программного обеспечения. Так, специальные программы могут просматривать весь состав многочисленных входных и выходных словарей, таблиц морфологических окончаний, синтактико-семантических кодов и других массивов информации с целью сплошной проверки, занесения новых признаков, изменения структуры данных.

Грамматика есть множество грамматических признаков данного языка, семантика — набор семантических признаков (или семантико-синтаксических, в той части синтаксиса, которая не охватывается грамматикой). Способ задания элементов в каждом множестве есть перечисление. Таким образом, составляющие предметного элемента **предметного компонента** суть образующие текст.

Переводной элемент предметного компонента образуется из типов переводных соответствий. Он содержит межъязыковые соответствия, также заданные в форме перечислений, одной из форм может быть табличная. Эквивалентные соответствия устанавливаются всегда простой операцией соотнесения: так, если слово всегда во всех контекстах переводится однообразно, то перед нами эквивалентное соответствие, например: *to inform* — *информировать*. Конечно, эквивалентность соответствий имеет место, как правило, только в рамках достаточно четко очерченных подязыков. В той мере, в какой каждое слово естественного языка может быть многозначно и соответственно по-разному переведено на другой язык, эквивалентные соответствия в некоторой абсолютной теории перевода не должны появляться в виде априорной сущности: эквивалентность в общей теории может быть только текстовой. Однако именно ограничение подязыков общения позволяет ввести эквивалентные соответствия, как это делается также и в переводческой практике, в качестве некоторой составляющей модели.

Вариантные соответствия суть такие, которые устанавливаются с помощью некоторого формализованного выбора. Примером здесь могут служить простые случаи контекстного разрешения многозначности отдельных слов. Так, существительное *personnel* может переводиться на русский язык либо как *личный состав*, либо *персонал*. В первом случае перед ним должны стоять определения *army* или *military*. Если таких определений нет, то слово означает *персонал*. Ясно также, что вариантные соответствия могут не охватить всего разнообразия переводных соответствий, кроме того, могут быть кон-

текстные ограничения. Тем не менее группа соответствий такого типа достаточно представительна, интуитивно допустима, безусловно существует в распоряжении переводчика и поэтому может быть выделена в составе модели.

Трансформационные соответствия самые сложные в системе соответствий данной языковой пары. Если эквивалентные и варианты соответствия довольно естественно задаются в словаре, то трансформационные в большей степени относятся к окказиональным. Первые две группы можно практически считать регулярными, а последнюю — окказиональной, если смотреть с общей точки зрения на текстовую эквивалентность. Однако в составляющую модели можно включить такие трансформационные соответствия, которые также имеют значительную долю регулярности. Например, передача английской пассивной конструкции с предлогом *by* русской активной, в которой английское существительное-деятель становится в русском подлежащим, субъект — объектом и происходит изменение формы глагола: *The book was read by the student* — *Студент прочитал книгу*. Преобразования такого рода безусловно можно считать регулярными и включать в некоторый список трансформаций, которые с точки зрения алгоритмической реализации производятся достаточно четко определенными алгоритмами.

Статистическая определимость сказывается на том, какие из трансформационных соответствий мы относим в разряд регулярных, а какие — окказиональных. Действительно, вышеприведенные примеры могут иметь разные статистические характеристики не только для разных языков, но и для разных подязыков внутри одного языка. Статистические данные явного вида, т. е. те, которые собраны на выборке из определенного объема текстов данной генеральной совокупности, могут не иметь убедительной достоверности для языковых единиц, больших, чем слово, однако могут быть получены надежно ориентирующие данные по некоторой совокупности лексических единиц, например, типа английской пассивной конструкции, которая может подсчитываться с помощью ассоциации с соответствующей лексикой (глагол *be*, предлог *by*). Пример такого подсчета может быть использован для определения частоты синтаксических конструкций.

Коллективный опыт составителей алгоритма позволит выделить в составе трансформационных преобразований регулярные или те, которые можно отнести по ряду соображений к таковым, и окказиональные. В этом деле статистические соображения могут играть вспомогательную роль. Если некоторое преобразование осуществляется достаточно просто алгоритмически, то его можно учесть в области регулярных, не предъявляя к нему достаточно строгих статистических критериев. Наоборот, если алгоритмическое преобразование делается сложно, то его целесообразно рассматривать в рамках иконического подхода, не стремясь создать алгоритмические процедуры.

В этом тоже важен коллективный опыт разработчиков алгоритма в целом.

Специфика перевода в рамках данной языковой пары также имеет большое значение. Здесь может меняться не только диапазон соответствий. Близость конструкций для одной пары может значительно влиять на набор и характер трансформационных преобразований, включаемых в список регулярных трансформаций. Опыт работы в системах АМПАР, НЕРПА и ФРАП показал, что список преобразований для трансформационных соответствий существенно зависит не только от выбранного подхода (так или иначе в рамках любого подхода необходимость соответствующих или эквивалентных трансформаций остается), но и от характера языковой пары. В немецком языке, например, потребности преобразования конструкций типа рамочных или с отделяемыми приставками значительно расширяют список преобразований, которые осуществляются на разных этапах анализа и синтеза [Марчук 1983].

Для определения состава трансформационных соответствий полезны данные, получаемые в результате сопоставительного анализа систем двух разных языков. Так, основой выделения списка трансформаций может служить метод анализа обратных переводов для заданных конструкций двух сравниваемых языков, который, например, был применен при сопоставлении видовременных значений грузинского и английского языков [Хизанишвили 1983]. В обратном переводе на материалах, полученных методом экспертных оценок от многих переводчиков, можно получить теоретическую модель типичной интерференции систем видовременных отношений обоих языков. Системный анализ глагольных парадигм обоих языков, указывающий на эту область, подтвердился экспериментально. При многократно повторенном переводе можно как бы кристаллизовать систему языка как особую категорию, относительно самостоятельную, и показать тенденцию интерференции систем. С точки зрения трансформационных соответствий можно выявить регулярность этих соответствий и найти наилучший и наиболее правильный путь их передачи, тем более если есть возможность создать теоретическую модель.

Несопоставимые категории, как показано в данном пособии, не вызывают особых трудностей при переводе. Это значит, что можно достаточно просто (речь идет о переводе и преподавании языков) подобрать регулярные соответствия тем случаям, которых нет в грамматике. Для несопоставимых категорий обоих языков сопоставительная грамматика обнаруживает дополненность обеих систем, и при этом, как видно из материала исследований, трудностей не возникает. В рамках данной пары языков можно провести типологию трудностей для каждой пары сопоставляемых явлений. Так, для рассмотренного соотношения были выявлены четыре случая. С помощью обратного перевода были найдены наиболее выгодные способы передачи соответствующего значения на тот или иной язык.

Предметный элемент предметного компонента образует исходную структуру компонента, на которую накладывается сеть, образованная переводным элементом — системой соответствий. Однако в целом предметный компонент отражает предмет, статическое описание сущности исходных данных, над которыми впоследствии производится операция перевода. Важный смысл предметного компонента, отличающего его от статического описания языковых данных, заключается в том, что это целевое описание языковых данных для перевода, а не вообще языковых данных, пригодных для чего угодно. Предметный компонент модели переводных соответствий есть специфически переводческое описание статики языковых данных, участвующих, необходимых и достаточных для перевода с данного естественного языка на другой, с учетом всех преобразований, которые могут для этого понадобиться.

Динамический компонент отвечает на вопрос: каким образом должен быть осуществлен перевод? Содержанием этого компонента является процесс установления соответствий, динамика нахождения переводных соответствий в тексте и конструирования некоторого выходного текста, находящегося со входным в отношении коммуникативной эквивалентности. Для осуществления процесса динамический компонент содержит особый инструментарий, отраженный в его структуре, и состоит из двух основных элементов: алгоритмического и программного. Структура динамического компонента изображена на рис. 21.

Алгоритмический элемент включает собственно алгоритм, трансляционную грамматику и словарь. Алгоритм, который подразделяется на общий, организующий систему в целом, и частные алгоритмы (этапов, частей этапов, отдельных операций анализа и синтеза), управляет использованием трансляционной грамматики и словаря на соответствующих стадиях нахождения соответствий и построения эквивалентного текста. Трансляционная грамматика есть специфическая бинарная грамматика, которая оптимальным для перевода образом организует грамматические признаки, нужные для

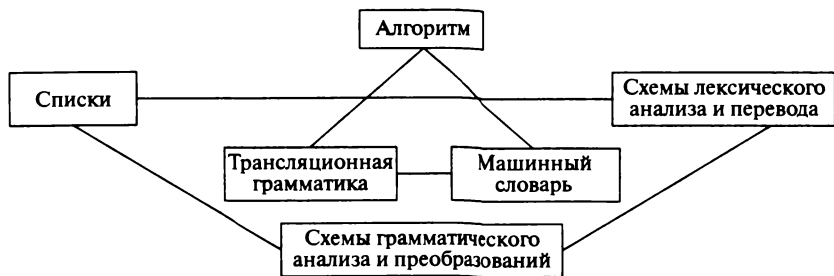


Рис. 21. Структура динамического компонента модели переводных соответствий

установления соответствий. Сюда входят признаки всех языковых уровней: морфологического, семантического, синтаксического (или семантико-синтаксического) и лексического. Словарь представляет собой упорядоченное для перевода собрание лексических единиц с необходимой информацией, которое организуется по трем основаниям: по признаку входного — выходного языка, однозначности — многозначности слов, фразеологических единств — отдельных лексем.

Поясним эту организацию словаря. В системе машинного перевода может быть много разнообразных словарей. Так, для перевода наиболее употребительных слов в рамках данной языковой пары может быть вынесен отдельно словарь наиболее частых слов с их переводами и/или алгоритмами разрешения их неоднозначности по тексту. Может быть, и в ряде случаев выделяется, словарь терминов, в который входят самые употребительные из образующих данную область знания. Эти словари могут также иметь разнообразную структуру; так, в словаре для перевода лексическим единицам входного языка могут быть сразу же сопоставлены единицы выходного, причем приписаны им еще до начала анализа. Словари выходного языка не привязаны непосредственно к входным единицам, но содержат всю информацию, необходимую для синтеза. Кроме того, информация к словарным единицам может делиться на постоянную и переменную части. Первая часть хранится в словаре и неизменно приписывается слову, вторая получается в результате работы алгоритмов анализа и синтеза и разнообразна для каждой лексической единицы.

Однозначность — многозначность лексических единиц чаще всего связана с наличием у каждой из них нескольких переводов. Многозначность, таким образом, приравнивается многопереводности. Для узких подязыков словари перевода многозначных слов могут быть достаточно ограниченными по объему, однако в любом случае они должны существовать, поскольку общеупотребительная и общетехническая лексика, неизбежно присутствующая в этих подязыках, для своего перевода требует соответствующего количества алгоритмов. Наконец, фразеологические единства, учитываемые в словаре, играют особую роль в том отношении, что описывающие их словари выполняют также и некоторую часть грамматического анализа: так, если сложному словосочетанию идиоматического характера, играющему роль, например, наречия (*as soon as*), приписывается признак «обстоятельства» или другая синтаксическая роль, то во всем дальнейшем анализе это словосочетание продолжает выполнять именно эту роль, облегчая дальнейший анализ. Поэтому словари фразеологических единств можно также в определенной степени назвать грамматическими.

Учитывая все сказанное, три приведенных выше деления словаря описывают практически основные релевантные для перевода характеристики словарной части.

Трансляционная грамматика содержит грамматические признаки входного языка, соотнесенные с грамматическими признаками язы-

ка выходного с целью дать систему их соответствий. Эта грамматика есть детализация и спецификация тех грамматических соотнесений, которые были рассмотрены в переводном элементе предметного компонента. Так, из указанного ранее соответствия грузинских и английских видовременных категорий, изученного при выделении «грамматики» в предметном компоненте, выделяются конкретные грамматические признаки, которые используются в трансляционной грамматике. Описание конкретной трансляционной грамматики английского языка в переводе на русский см. [Марчук 1983].

Алгоритм, выступающий в данной модели в качестве составляющей алгоритмического элемента, есть сугубо лингвистическое представление. Он определяет общий лингвистический замысел системы, порядок анализа, синтеза и перевода в данной последовательности действий, в нем используются лингвистические сведения разных языковых уровней. На этом этапе также важен выбор алгоритмического языка, не с точки зрения математической модели последовательности действий, а с позиции удобства содержательных запросов текста и лингвистических баз данных для этапов обработки словоформ и текста в целом. Программный язык записи действий алгоритма должен быть достаточно простым и удобным для лингвиста.

Программный элемент представляет собой совокупность программных решений для этапов алгоритма. Название «программный» показывает, что здесь происходит выход на уровень реализации. Становится важным язык программирования, алгоритмический язык, система реализации лингвистического алгоритма в виде последовательности этапов. Так, учитываются возможности, которые предоставляет лингвисту та или иная система программирования. Составляющие программного элемента, а именно списки и схемы, являются функцией от принятого способа программного представления частных алгоритмов анализа и синтеза. Списки являются простейшей составляющей, наименее зависящей от типа языка. Схема есть программная реализация лингвистического алгоритма в стандартных операторах, каждый из которых представляет собой тот или иной способ реализации лингвистической операции достаточной сложности: например, поиск слова в тексте слева или справа от заданного слова осуществляется одним стандартным оператором. Следует учесть, что при таком поиске должны пропускаться словоформы и другие элементы текста, не относящиеся к содержанию конкретного задания.

В анализирующей части составляющие программного элемента могут рассматриваться по аналогии с типом переводных соответствий: так, списки реализуют соответствия эквивалентного типа, схемы лексического анализа — вариантного типа, грамматического — трансформационные преобразования. Схемы разных типов характеризуются разной степенью сложности используемых программных операторов и самим их числом.

Контекстное разрешение лексической многозначности в машинном переводе

Принцип построения контекстологического словаря

В системе машинного перевода АМПАР был применен автоматический контекстологический словарь. В машинном переводе, в том числе и в современных системах, такого рода словари, как видно из результатов перевода, используются достаточно редко, поэтому число ошибок в переводе многозначных слов достаточно велико. Разработанный и опубликованный контекстологический словарь явился результатом труда большого коллектива разработчиков в течение нескольких лет. Большая трудоемкость составления такого словаря, по-видимому, и является основной причиной того, что подобные инструменты повышения эффективности машинного перевода достаточно редки в действующих системах.

Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский имеет целью обеспечить перевод многозначных слов приведенного в нем списка с учетом контекстного окружения этих слов по строго формальным признакам с помощью алгоритмической процедуры. Основное назначение словаря — использование в системах автоматической обработки текстов [Марчук 1976].

Отличие этого словаря от традиционных переводных состоит в том, что он составлен:

- для сравнительно ограниченной, хотя и достаточно широкой, области современного английского языка — для подязыка газетных сообщений на политические, публицистические, экономические и научно-технические темы;

- не на основе существующих переводных и толковых словарей, а на результатах обследования текстов указанного характера объемом около 700 000 словоупотреблений. В качестве исходного материала для его составления служили специальные словари-конкордансы, в которых каждое слово, каждая словоформа дается в том контексте, в котором оно употреблено в исходном массиве текстов. На базе этих словарей переводчик-составитель контекстологического словаря определял однозначность или многозначность нужного слова в зависимости от контекста. При этом многозначность приравнивалась к «многоязычности» — если слово имело разные переводы, то оно признавалось многозначным. При наличии многозначности переводчик определял, какие составляющие лингвистического контекста разрешают данную конкретную многозначность. Далее он строил алгоритм запроса контекста на наличие в нем детерминант, т.е. признаков, разрешающих многозначность. Контекстологический словарь является собранием таких алгоритмов для всех отмеченных в исходном объеме текстов многозначных слов;

– осуществляемый с его помощью перевод, не имеющий диагностирующих форм, реализуется в виде так называемого «общего выхода» алгоритма, который не определяется дистрибуцией, а выбирается по статистическим характеристикам.

Контекстологический словарь содержит большое количество словосочетаний разной степени устойчивости. Идиомы в собственном смысле в него не вошли, так как они образуют особый словарь оборотов, который в системе машинного перевода работает первым. Однако вследствие известной трудности отделения идиом от свободных словосочетаний некоторое количество лексических единиц, представляющих фразеологические словосочетания, все же может быть найдено в контекстологическом словаре.

Применение контекстологического словаря

Структура и содержание контекстологического словаря делают возможными его разнообразные применения, главными из которых являются:

– непосредственное использование в системах обработки английских текстов, машинного перевода или информационного поиска для разрешения многозначности ключевых слов или перевода многозначных слов на русский язык. При этом нужно только выбрать форму кодирования словаря. Исходная грамматическая информация, на которую опирается в своей работе словарь, и алгоритм, в рамках которого он работает, могут быть при необходимости сравнительно просто модифицированы;

– применение в качестве справочного пособия при обучении английскому языку, поскольку словарь отражает наиболее актуальную сочетаемость английских слов, а также в качестве пособия переводчикам по публицистической (газетной) тематике;

– привлечение в качестве основы для программированного обучения английскому языку. Для такого использования удобна нынешняя его форма, предназначенная для машинного перевода;

– использование для научно-исследовательской работы лингвистов в области современной английской лексикологии, лексикографии и машинной обработки текстов, а также для других целей.

Строение контекстологического словаря и порядок пользования словарем

Теоретической основой контекстологического словаря является теория детерминант [Марчук 1973]. Согласно этой теории, каждое значение (перевод) многозначного слова, отличающееся от наиболее общего значения («общего выхода» алгоритма перевода), детерминируется в контексте другими словами, с которыми данное слово сочетается, или определенными грамматическими признаками. Эти

слова, а также классы или группы слов и грамматические признаки называются детерминантами.

Контекстологический словарь, представляя собой собрание алгоритмов, базируется на грамматическом описании английского и русского языков и предусматривает определенные требования к общему или дополнительным алгоритмам, совместно с которыми работает данный словарь. Исходная грамматика описывается следующими основными положениями:

- все слова английского языка разделены на лексико-грамматические классы (части речи): существительное, прилагательное, глагол, наречие, предлог, союз, знак препинания. Местоимения и числительные разделены соответственно их функциям между существительными и прилагательными, образуя подклассы. В особый класс выделены цифры. В классах существительных, прилагательных и глаголов выделены по семантическому и дистрибутивному признакам подклассы, например: «одушевленные существительные», «качественные и относительные прилагательные» и т. п. Знаки препинания выделены в особый класс вследствие их роли в анализе и синтезе предложения;

- лексико-грамматическая омонимия (омография) — принадлежность словоформы к тому или иному лексико-грамматическому классу — разрешается до работы контекстологического словаря специальными алгоритмами, составляющими этап разрешения омографии;

- словарь входного (английского) языка является словарем основ. Поэтому в качестве лексических детерминант выступают не словоформы, а слова-основы. Морфологический анализ предусматривает выделение в составе словоформ обычных для английского языка суффиксов и окончаний. Морфологические признаки слова также могут выступать в качестве детерминант.

В системе алгоритмов, в которой работает данный контекстологический словарь, предусмотрены следующие операции:

- в пассивной конструкции правая лексическая детерминанта в алгоритмах перевода глаголов становится левой — меняет направление;

- прилагательные-определения образуются от однокорневых существительных;

- поиск лексических детерминант осуществляется с пропуском несущественных для данного случая слов согласно специальным подпрограммам пропуска.

Алгоритм перевода многозначного слова представляет собой последовательность запроса контекста на наличие в нем детерминант, каждая из которых однозначно определяет перевод этого слова. Операции по анализу контекста и прочие действия, связанные с переводом, выполняются стандартными операторами, каждый из которых представляет собой стандартную подпрограмму компьютера.

В контекстологическом словаре используются следующие стандартные операторы:

– поиска в тексте лексически выраженных детерминант (отдельных слов, списков слов или лексико-грамматических классов или подклассов слов). В этом операторе задаются: вид поиска (слово, список или класс слов), направление поиска (вправо или влево от основного слова), правила пропуска несущественных для данного вида поиска слов;

– проверки слова (или лексической детерминанты) на наличие определенного морфологического признака, такого, как «множественное число», «суффикс», и грамматического признака (время глагола) и т. п.;

– приписывания перевода. Перевод может быть приписан либо только основному слову, либо также и лексической детерминанте.

Алгоритм перевода многозначного слова представляет собой перенумерованную последовательность операторов, выполняемых в том порядке, в котором они записаны. При ответе «нет» операторов поиска или проверки происходит отход к тому оператору, номер которого указан в операторе проверки или поиска. Если лексически выраженные детерминанты образуют списки, эти списки приводятся в конце словарной статьи (алгоритма), а в самом операторе обозначается номер списка.

Операторы обозначаются следующими терминами: *rgt* — проверка вправо от основного слова; *lft* — проверка влево, *tr* — оператор перевода, *ch* — оператор проверки на грамматический признак, *list* — обозначение списка слов, на которые производится проверка.

Приведем пример алгоритма перевода глагола английского *back*:

BACK (глагол)

1	Rgt	3	Down	3
2	Tr		Оступиться	
3	Rgt		Out	6
4	Rgt		Of	6
5	Tr	3, 4	Уклониться от	
6	Rgt	3	Away	8
7	Tr		Отказаться	
8	Rgt		Up	10
9	Tr	8	Поддержать	
10	Rgt	1	Существительное	13
11	Rgt		Up	13
12	Tr	11	Поддержать	
13	Tr		Поддержать	

Первый столбец алгоритма — номер оператора по порядку. Второй столбец — номер подпрограммы пропуска несущественных слов (прилагательных, определений). Третий столбец — русский перевод. Четвертый столбец — отход при ответе «нет» к следующему опера-

тору. Последняя строчка — общий выход алгоритма. Если в тексте нет ни одной из предыдущих детерминант, дается перевод, указанный в общем выходе.

В заключение можно сформулировать основные характеристики контекстологического словаря. Это, во-первых, базирование его на исходном корпусе текстов. Чем представительнее корпус, тем более эффективен перевод многозначных слов с помощью такого словаря. Во-вторых, существуют предварительные условия работы такого словаря в автоматизированной системе, в частности должна быть разрешена лексико-грамматическая омонимия. В-третьих, должен быть проведен морфологический и грамматический анализ, с тем чтобы стало возможным использование грамматических детерминант — окончаний, суффиксов и т. п. Наконец, можно отметить достаточную трудоемкость разработки такого словаря, поскольку он требует предварительного создания конкордансов.

Тем не менее только такой словарь способен разрешить важнейшую для автоматической обработки текстов проблему многозначности лексических единиц, которая лежит в основе возможности передачи смысла текстов.

Контрольные вопросы

1. Имеет ли отношение формализация процесса перевода к теории языка?
2. Каковы связь и взаимовлияние проблематики искусственного интеллекта и автоматизации перевода?
3. Как можно описать структуру модели перевода по принципу переводных соответствий?
4. Какова последовательность реализации модели перевода по принципу переводных соответствий в конкретной языковой паре?
5. Каково содержание понятия «единица перевода»?
6. В чем заключается роль контекста в разрешении лексической многозначности?
7. В чем заключается принцип построения контекстологического словаря?

ЗАКЛЮЧЕНИЕ

В данном пособии мы проанализировали основные модели как «человеческого», так и машинного перевода. Следует отметить, что основная часть этих моделей была представлена достаточно давно и их существование и последующее развитие доказывают, что они реально отражают соответствующие проблемы изучения такого сложного интеллектуального процесса, каким является перевод с одного языка на другой. Вместе с тем, каждая модель отражает какие-то определенные аспекты перевода, один аспект или несколько. Связь перевода с другими языковыми закономерностями, с различными уровнями языка, с экстралингвистической реальностью безусловно отражается в моделировании человеческого (естественного) интеллекта. В этой связи мы согласны с точкой зрения академика Ю. В. Рождественского о том, что центральной проблемой искусственного интеллекта является машинный перевод.

По нашему мнению, вершиной моделирования перевода, его высшим достижением является высококачественный машинный перевод. Возможно, многие лингвисты, воспитанные в традиционном гуманитарном лингвистическом духе, отрицательно отнесутся к тому, что «механическое препарирование языка в промышленных целях», как иногда определяют машинный перевод, объявляется высшим достижением моделирования перевода.

Лингвистическому, филологическому, толковательному, литературному описанию перевода как предмета и объекта исследования посвящена обширная литература, и число работ на эту тему продолжает увеличиваться. В настоящий обзор включена их относительно малая часть, хотя и освещено максимальное число наиболее выдающихся исследований по этой теме. Однако часто многие из этих моделей не дают возможности или не рассматривают построение «воспроизводящих» лингвистических моделей перевода, что автор пытался сделать центральным в своем рассмотрении.

Машинный перевод представляет собой наиболее сложный и вместе с тем наиболее актуальный вид переводческой деятельности, будучи центральной проблемой искусственного интеллекта, являющегося неотъемлемой частью перевода в новых технологических условиях, на новом этапе семиозиса. Машинный перевод объясняет и формализует скрытые механизмы этого вида деятельности человека. Для того чтобы научить компьютер «понимать» в рамках

перевода достаточно сложные тексты, надо составить и запрограммировать алгоритмы морфологического, синтаксического, лексического, семантического анализа, опробовать их на текстах, убедиться, что каждый из них решает свою задачу в намеченных пределах точности, скомпоновать все эти алгоритмы и перевести с их помощью новый текст, на базе которого они не составлялись. Этим уже много лет занимаются большие и малые группы исследователей во многих странах мира, и в этой области достигнут определенный прогресс, во всяком случае, есть действующие системы машинного перевода, которые удовлетворительно переводят не слишком сложные тексты.

Следует отметить, что автор настоящего пособия является сторонником определенного направления в машинном переводе, за много лет своего развития создавшего немало разных научных школ, придерживающихся различных взглядов на то, как надо «правильно» решать эту проблему. Будучи сторонником такого моделирования машинного перевода, которое уделяет основное внимание построению этапа переводных соответствий, модели «текст-текст» и не стоит на позиции «машинный перевод без перевода, без машин, без алгоритмов», автор этого пособия также старался быть объективным в описании и оценке проблематики направления «текст-смысл-текст». Кстати сказать, работы именно этого направления достаточно широко представлены и описаны в научной литературе по проблеме автоматизации перевода в отличие от «воспроизводящих» моделей, которые часто, особенно в настоящее время, представляют технологические секреты фирм и не описываются подробно. Однако в оценке эффективности этих моделей помогает принцип «черного ящика», когда по результатам перевода можно судить о принципиальном устройстве модели. В этих эмпирических системах часто сам принцип и не очень осознавался. Осознание принципа действующих систем поможет его развитию, если он полезен, или его модификации, если он не очень хорош.

Методы формального моделирования перевода к настоящему времени разработаны в такой степени, что они дают возможность с заданной степенью точности построить алгоритмы анализа, синтеза, необходимых преобразований структур, и создать словари, позволяющие организовать машинный перевод определенного вида текстов с соблюдением заданных технико-экономических параметров. Таков наиболее общий вывод, который, по мнению автора, можно сделать из рассмотрения существующего положения с моделированием перевода. Однако детализация этого положения довольно значительна. Модели перевода, известные в настоящее время, не дают возможности получать машинный перевод такого же качества, каким бы его производил человек, без значительных усилий пост-, интер- или пред-редактора. При этом мы отвлекаемся от того, что и человеку-переводчику нужен редактор и что теоретически полностью адекват-

ного и эквивалентного перевода любого текста с одного естественного языка на другой не существует. Всякий перевод есть всегда достаточно точное приближение к оригиналу, но не сам оригинал.

Машинный перевод в огромной степени повлиял на традиционную теорию перевода благодаря главным образом концепции воспроизводимости моделей перевода. Он также способствовал возникновению многих новых направлений не только в прикладной и компьютерной лингвистике, но и в языкознании в целом, расширив понятие моделей и моделирования. Что касается собственно перевода, то автор согласен с точкой зрения академика Ю. В. Рождественского о том, что наилучшим вариантом является перевод буквальный, возможный только при полном равенстве культур, технологий, научных достижений. Многие переводчики хорошо помнят то время, когда в научно-техническую литературу на русском языке входили понятия *файл*, *процессор*, *инфраструктура* и пр. Дело здесь не в обычном калькировании, в результате которого, например, в наш обиход входят слова типа *напряженность*, *позиционировать*, *тренд* и т. п. или обороты типа *американское военное присутствие*. Дело в уравнивании культур и соответствующих языковых картин мира, в сопоставимых уровнях описания, которые были бы понятны. Общечеловеческие ценности остаются неизменными на века, почему мы и можем понимать Шекспира, однако трудно себе представить, как мог бы понять народ, не знакомый с вычислительной техникой, описание вычислительного процесса на компьютере. Поэтому «буквальный», т. е. наиболее близкий к тексту, не нуждающийся в толкованиях, описательных конструкциях, объяснениях, переносе значений и т. п., перевод действительно был бы идеальным. В этом смысле автомат лучше всех делал бы такой перевод.

Другим вопросом общетеоретического характера, возникающим при моделировании перевода, является соотношение личного и безличного в переводе. Переводчик не должен отражать свои личные взгляды и чувства. Однако он иногда это делает — брать, например, стилистические характеристики научно-технического перевода в исследовании Л. И. Борисовой. Это отражается в предпочтении определенной лексики, в выборе синтаксических конструкций, не говоря уже об ошибках перевода вследствие неполноты понимания. Это так или иначе характеризует работу человека-переводчика, с этим хорошо знакомы редакторы. Машина-переводчик, делая гораздо больше ошибок, чрезвычайно затрудняющих иногда работу редактора, тем не менее максимально объективна в своих переводах и не думает скрывать своих ошибок. Это является следствием того, что алгоритмы перевода суть следствие коллективных усилий составителей схем анализа, синтеза, словарей системы перевода. «Коллективный» перевод компьютера есть некоторый максимально обезличенный «идеал», к которому должен стремиться каждый перевод независимо от того, перевод ли это поэзии или технической документации.

Машинный перевод в существенной степени связан с типологией перевода. Последняя относительно мало интересовала теоретиков перевода до настоящего времени. Помимо упоминания «особых» видов перевода, таких, как перевод документов, научно-технический перевод, перевод поэзии, художественный перевод и пр., более точной типологии, например, в аспекте подъязыков практически не существовало, поскольку она не очень интересовала человека-переводчика, способного передать практически любой текст. Машинный перевод, особенно его практическое использование, заставил вновь поставить вопрос о типологии перевода не только в практических коннотациях, но и в теории — что, зачем и как мы переводим. Образность, столь трудная в поэтическом и художественном переводе, запрещена в переводе техническом. Машинный перевод заставляет пристально посмотреть на особенности «особых» видов перевода — юридических, дипломатических, политических и прочих документов, заставляет создать и описать инвентарь языковых средств, определяющих соответствующие «особенности» этих видов перевода. Деление перевода на собственно перевод — на уровне языковых средств — и интерпретацию, когда переводчик передает содержание текста «своими словами», деление, которое далеко не однозначно и в традиционном чисто филологическом истолковании, получает определенное развитие в дальнейших исследованиях по теории, практике и типологии перевода.

Особое место занимают понятия «единица перевода» и «переводное соответствие». Опыт практического осуществления «человеческого» перевода, эксперименты и практика перевода машинного, попытки объяснить механизмы перевода приводят к выводу о полезности такого лингвистического конструкта, каким является единица перевода. Она различна и, вероятнее всего, будет разной для разных типов перевода, однако само ее существование и стремление выделить ее со стороны переводчика (и редактора) дают определенный полезный и конструктивный вклад в теорию и практику. Поэтому признание существования такой единицы есть положительный фактор.

Переводные соответствия для данной конкретной языковой пары и составляют тот конечный, единственный и решающий арсенал средств, с помощью которых осуществляется перевод. «Буквальный» перевод, с помощью которого можно наиболее точно передать форму и содержание исходного текста, в соответствии со сказанным выше становится возможным лишь при наличии полного списка соответствий и набора языковых средств, обеспечивающих удовлетворительную передачу содержания. Структура системы соответствий может быть разной, она зависит от типологии соответствий, которая, в свою очередь, определяется принятой моделью перевода. В машинном переводе роль соответствий в настоящее время общепризнанна, хотя в течение многих лет некоторые специалисты пытались уйти от этой проблемы, возлагая задачу привлечения языковых средств на алго-

ритм перевода, составленный математиками. Этап трансфера в современных действующих и проектируемых системах машинного перевода является отражением ведущей роли переводных соответствий в моделировании перевода для компьютера и в «самостоятельном» моделировании, если такое возможно.

Все предыдущее, особенно представление о единице перевода, роли переводных соответствий, типологии перевода, зависимости видов перевода и вообще перевода от типа текстов, объясняет большую роль понятия подязыка для моделирования перевода. Единый национальный язык как средство коммуникации может быть в практических целях рассмотрен как состоящий из разных подязыков, различным образом связанных между собой. Успех моделирования перевода зависит от того, каковы структура, лексика, грамматика, синтаксис, семантика, коммуникативные характеристики того или иного подязыка. Точнее, на основе этих данных должно быть принято решение относительно того, насколько возможно построить практическую модель автоматического перевода текстов данного подязыка. Практические потребности такого перевода, успех методов, ориентированных на ограниченную языковую специфику отдельных сфер языкового общения, заставили подробнее рассмотреть языковое строение подязыков. В настоящем пособии не очень подробно освещен вопрос о сущности и строении подязыков, но указана на важность этой проблемы для эффективного моделирования перевода.

Можно сказать, что машинный перевод как некоторая наиболее сложная форма моделирования перевода более тесно связан с информационной деятельностью, чем перевод традиционный. Последний существует самостоятельно, обладая большой спецификой по сравнению со всеми другими видами информационной деятельности. Машинный перевод не отделим от информационной деятельности других видов: от реферирования, индексирования, сигнального информирования, быстрой и точной обработки документов. Представлять машинный перевод как отдельный и самостоятельный вид информационной работы, безусловно, можно, однако это означает отрыв от экономических, технологических и социальных факторов, во взаимодействии с которыми развивается перевод. Это означает, что моделирование для машинного перевода есть задание комплексное. В этом пособии предпринята попытка показать связь моделей машинного перевода с моделированием мышления, интеллекта, с другими информационными задачами.

Подводя общий итог рассмотренных проблем, можно сказать, что моделирование перевода как сложной интеллектуальной деятельности человека будет продолжаться с разных теоретических и практических позиций, принося новые открытия и концепции, что, безусловно, будет полезно как для прикладного, так и для общего и сравнительно-сопоставительного языкознания в сегодняшнем активно развивающемся мире.

СПИСОК ЛИТЕРАТУРЫ

Актуальные проблемы лингвистики и лингводидактики иностранного языка делового и профессионального общения. III Международная научно-практическая конференция: материалы конференции. — М., 2008.

Андреева Е. С. Диалектика текста. Опыт логико-лингвистического синтеза. — М., 2001.

Бархударов Л. С. Язык и перевод. — М., 1975.

Бельская И. К. Язык человека и машина. — М., 1969.

Берков В. П. Вопросы двуязычной лексикографии. — Л., 1973.

Борисова Л. И. Лексические особенности англо-русского научно-технического перевода. — М., 2005.

Вайнцвайг М. Н., Полякова М. П. Механизм мышления и моделирование его работы в реальном времени. // Интеллектуальные проблемы и их моделирование. — М., 1987.

Валипур Али Реза. Анализ и синтез глагольных форм и конструкций при машинном переводе с русского языка на персидский язык: канд. дис. — М., 1998.

Виссон Линн. Синхронный перевод с русского на английский. — М., 1999.

Влахов С., Флорин С. Непереводимое в переводе. — М., 1986.

Головин Б. Н. Введение в языкознание. — М., 1966.

Гринев-Гриневич С. В. Терминоведение. — М., 2008.

Гринев-Гриневич С. В., Сорокина Э. А., Скопюк Т. Г. Основы антропологической лингвистики (к лексическим основаниям эволюции мышления человека): учеб. пособие. — М., 2005.

Денисов П. Н. Лексика русского языка и принципы ее описания. — М., 1993.

Денисов П. Н. Очерки по русской лексикологии и учебной лексикографии. — М., 1974.

Денисов П. Н. Принципы моделирования языка. — М., 1965.

Ефимов Н. Н., Фролов В. С. Основы информатики. Введение в искусственный интеллект. — М., 1991.

Заслонко А. Н., Хахалин Г. К. Лингвистический транслятор для системы машинного перевода // Международный семинар по машинному переводу. — Тбилиси, 1989.

Звегинцев В. А. Предложение и его отношение к языку и речи. — М., 1976.

Зелко В. М. Проблемы разработки лингвистического обеспечения системы китайско-русского информационного машинного перевода: канд. дис. — М., 1991.

- Зубова И. И.* Информационные технологии в лингвистике. — Минск, 2001.
- Кан Д.* Взломщики кодов. — М., 2000.
- Караулов Ю. Н.* Общая и русская идеография. — М., 1976.
- Колшанский Г. В.* Контекстная семантика. — М., 1980.
- Комиссаров В. Н.* Современное переводоведение. — М., 2004.
- Комлев Н. Г.* Слово в речи. Денотативные аспекты. — М., 1992.
- Корнилов О. А.* Языковые картины мира как производные национальных менталитетов. — М., 2003.
- Коростелев Л. Ю.* Стохастический подход к графематике слов (проблема диагностики и коррекции искажений в системе машинного перевода): канд. дис. — М., 1985.
- Котов Р. Г., Новиков А. И., Скокан Ю. П.* Прикладная лингвистика и информационные технологии. — М., 1987.
- Котов Р. Г., Марчук Ю. Н., Нелюбин Л. Л.* Машинный перевод в начале 80-х годов. — М., 1983.
- Кравченко С. Ю.* Формализация процесса лексико-грамматического анализа текстов // Актуальные проблемы прикладной лингвистики. — Минск, 2008. — Ч. 2.
- Кривоносов А. Т.* Система классов слов как отражение структуры языкового сознания. — М.; Нью-Йорк, 2001.
- Кривоносов А. Т.* Язык, логика, мышление. Умозаключение в естественном языке. — М.; Нью-Йорк, 1996.
- Лингвистический энциклопедический словарь. — М., 1990.
- Линс У.* Опасный язык: книга о преследованиях эсперанто. — М., 1999.
- Любимова А. А.* Интерпретация языкового манипулирования в контексте проблем адекватного перевода // Теория и практика перевода. — М., 2005.
- Магомедова А. Н.* Корпусная лингвистика и контекстное разрешение лексической многозначности слов: канд. дис. — Нальчик, 2004.
- Мартыненко Г. Я.* Некоторые методологические проблемы корпусной лингвистики. Материалы III Международной научной конференции «Прикладная лингвистика в науке и образовании». — СПб., 2006.
- Марчук Ю. Н.* Язык делового общения и лексическая многозначность. Материалы III Международной конференции «Актуальные проблемы лингвистики и лингводидактики иностранного языка делового и профессионального общения. — М., 2008.
- Марчук Ю. Н.* Машинный перевод с русского языка на восточные языки (некоторые аспекты) // Перевод: традиции и современные технологии. — М., 2002.
- Марчук Ю. Н.* Модель «текст-текст» и переводные соответствия в теории машинного перевода // Проблемы компьютерной лингвистики. — Минск, 1997.
- Марчук Ю. Н.* Теория и практика машинного перевода. // Русский филологический вестник. — М., 1996.
- Марчук Ю. Н.* Основы терминографии. — М., 1992.
- Марчук Ю. Н.* Методы моделирования перевода. — М., 1985.
- Марчук Ю. Н.* Проблемы машинного перевода. — М., 1983.

- Марчук Ю. Н.* Вычислительная лексикография. — М., 1976.
- Марчук Ю. Н.* Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский. — Ч. 1: Существительные, прилагательные; Ч. 2: Глаголы, наречия. — М., 1976.
- Марчук Ю. Н.* Опыт машинной реализации дистрибутивной методики определения лексических значений // Статистика речи и автоматический анализ текста. — Л., 1973.
- Марчук Ю. Н., Моторин Ю. А.* Основные принципы автоматизации перевода с английского языка на русский. Вопросы радиоэлектроники. — М., 1970. — Вып. 7. Серия ЭВТ.
- Маслов Ю. С.* Введение в языкознание. — М., 1987.
- Мельчук И. А., Равич Р. Д.* Автоматический перевод. 1949—1963. — М., 1967.
- Миньяр-Белоручев Р. К.* Как стать переводчиком. — М., 1994.
- Мохаммади Мохаммад Реза.* Система русско-персидского машинного перевода на основе переводных соответствий: канд. дис. — М., 1998.
- Нелюбин Л. Л.* Введение в технику перевода. — М., 2007.
- Нелюбин Л. Л.* Толковый переводоведческий словарь. — М., 2006.
- Новиков А. И.* Доминантность и транспозиция в процессе осмысления текста // Проблемы прикладной лингвистики. 2001. — М., 2002.
- Новиков В. А.* Трансфер в современных системах машинного перевода: канд. дис. — М., 2001.
- Панов Д. Ю., Ляпунов А. А., Мухин И. С.* Автоматизация перевода с одного языка на другой. — М., 1956.
- Петров В. В., Переверзев В. Н.* Обработка языка и логика предикатов. — Новосибирск, 1993.
- Пиотровский Р. Г.* Лингвистическая синергетика: исходные положения, первые результаты, перспективы. — СПб., 2006.
- Пиотровский Р. Г.* Лингвистический автомат и его речемыслительное обоснование. — Минск, 1999.
- Пиотровский Р. Г.* Инженерная лингвистика и теория языка. — Л., 1979.
- Пиотровский Р. Г.* Текст, машина, человек. — Л., 1975.
- Поспелов Д. А.* О «человеческих» рассуждениях в интеллектуальных системах. Логика рассуждений и ее моделирование // Вопросы кибернетики. — М., 1983.
- Потапова Р. К.* Новые информационные технологии и лингвистика. — М., 2002.
- Потапова Р. К.* Речь: коммуникация, информация, кибернетика. — М., 2001.
- Почепцов Г. Г.* Теория коммуникации. — М., 2003.
- Рахимбердиев Б. Н.* Эволюция семантики экономической терминологии русского языка в XX веке: канд. дис. — М., 2003.
- Ревзин И. И., Розенцвейг В. Ю.* Основы общего и машинного перевода. — М., 1964.
- Рецкер Я. И.* Теория перевода и переводческая практика. — М., 1974.
- Рождественский Ю. В.* Философия языка. Культуроведение и дидактика. Современные проблемы науки о языке. — М., 2003.

- Рождественский Ю. В., Волков А. А., Марчук Ю. Н.* Введение в прикладную филологию. — М., 1988.
- Рождественский Ю. В.* Введение в общую филологию. — М., 1979.
- Рябцева Н. К.* Язык и естественный интеллект. — М., 2005.
- Рябцева Н. К.* Язык и естественный интеллект: конгруэнтность, асимметрия, дополнительность // Проблемы прикладной лингвистики. — М., 2004.
- Рябцева Н. К.* Лингвистическое моделирование естественного интеллекта и представление знаний // Проблемы прикладной лингвистики. — М., 2002.
- Садохин А. П.* Теория и практика межкультурной коммуникации. — М., 2004.
- Светова С. Ю., Косматова Е. В.* Системы автоматизированного перевода PROMPT. Системы Translation Memory Trados. Интеграция Trados и PROMPT // Перевод: традиции и современные технологии. — М., 2002.
- Советский энциклопедический словарь. — М., 1980.
- Сотников Д. И., Сотников А. Д.* Язык международный рациональный арифметический (Lingvointeraciaraarimetika (LIRA)). — М., 2004.
- Тайеби Мосавимиянгах.* Морфологический анализ в системе англо-персидского машинного перевода: канд. дис. — М., 2002.
- Танасиенко А. В.* О некоторых методологических принципах построения систем машинного перевода нетрадиционного типа, базирующихся на обобщенно-модельной концепции лингвистического знака // Международный семинар по машинному переводу. — Тбилиси, 1989.
- Уивер У.* Перевод // Машинный перевод. — М., 1957.
- Татаринов В. А.* Общее терминоведение: энциклопедический словарь. — М., 2006.
- Тулдава Ю.* Проблемы и методы квантитативно-системного исследования лексики. — Таллин, 1987.
- Хизанишвили Н. Г.* Сопоставление частных грамматических категорий английского и грузинского глаголов: канд. дис. — М., 1983.
- Хроменков П. Н.* Современные системы машинного перевода. — М., 2005.
- Чжань Дэхуа.* Лексическая безэквивалентность русского языка по отношению к китайскому языку. (на материале переводов в русско-китайском словаре): канд. дис. — М., 1994.
- Шабанов-Кушнарченко Ю. П.* Теория интеллекта. Математические средства. — Харьков, 1984.
- Швейцер А. Д.* Перевод и лингвистика. — М., 1973.
- Шемакин Ю. И.* Введение в информатику. — М., 1985.
- Тезаурус научно-технических терминов / под. ред. Ю. И. Шемакина. — М., 1972.
- Эммарлу Рамезанали.* Синтаксический анализ и синтез именных словосочетаний при машинном переводе с русского языка на персидский: канд. дис. — М., 1998.
- Язык и машины (применение электронных вычислительных машин в переводе и в исследовании языка): отчет Наблюдательного комитета по автоматической обработке текстов. — Национальная Академия наук США, На-

циональный научно-исследовательский совет). — М.: журнал Научно-техническая информация. ВИНТИ АН СССР. Информационные процессы и системы. Сер. 2. — 1968. — № 8.

Better Translation for Better Communication. G. Van Slype, J. F. Guinet, Seitz, S. Benejam. — Pergamon Press, Oxford etc, 1983.

Boitet Ch. Twelve Problems for Machine Translation. International Conference on Current Issues in Computational Linguistics. — Universiti Sains Malaysia, Penang, Malaysia, 1991, Proceedings.

Early Years in Machine Translation. Memoirs and biographies of pioneers. — Ed. by W. John Hutchins. John Benjamins Publ. Co., Amsterdam/Philadelphia, 2000.

Gaussier E., Hull D., Ait-Mokhar S. Term Alignment in Use: Machine-aided Human Translation. — In: Parallel Text Processing. — Ed. by J. Veronis. Dordrecht, Netherlands, 2000.

Hill J. C. Language Acquisition // Encyclopedia of Artificial Intelligence. — New-York, E. A., 1987.

Kahn D. The Codebreakers. — New York: MacMillan, 1968.

Kirschner Z. A Dependency-based analysis of English for the purpose of machine translation. — In: Explizite Beschreibung der Sprache und automatische Textbearbeitung. Pr., 1982, Bd. IX, S 157.

Miram G. E. Translation Algorithms. — Kyiv: Twin Inter, 1998.

Prospects for a New Structuralism. — Univ. of Ottawa. Amsterdam-Philadelphia — Benjamins, 1992.

Tsujii J. Machine Translation: Productivity and Conventionality of Language. // Current Issues in Linguistic Theory — 136. Ed. by R. Mitkov et al. John Benjamins Publ. Co., Amsterdam-Philadelphia, 1997.

Tsutsumi J., Nitta T., Ono K., Nobesawa Sh., Nakanishi M. Multilingual Machine Translation Based on Statistical Information. QUALICO-94, 2nd Intern. Conf., on Quantitative Linguistics. Proceedings. — Moscow, 1994.

Vazquez-Ayora G. On the Notion of an analytical Unit of Translation. — Babel, 1982, vol. XXVII, No. 2.

Wiggers Pascal, Rothkrantz Leon J. M. Exploring the Influence of Speaker Characteristics on Word Use in a Corpus of Spoken Language Using a Data Mining Approach. SPECOM-2007. — Moscow, 2007.

ОГЛАВЛЕНИЕ

Предисловие.....	3
РАЗДЕЛ I. СОВРЕМЕННЫЕ ПРОБЛЕМЫ ЯЗЫКОВОГО ПЕРЕВОДА	7
Глава 1. Место языкового перевода в современной коммуникации	7
Преодоление языковых барьеров.....	7
Перевод и культура	9
Перевод и лингвистика.....	11
Понятие модели в лингвистике.....	14
Типология перевода	15
Массовая коммуникация — новый этап развития семиозиса	19
Глава 2. Универсальные языки и их роль в массовой коммуникации... 22	22
Универсальный язык как средство общения	22
Универсальный язык как фундамент логики общения.....	23
Универсальный язык как основа автоматизации перевода	24
Глава 3. Терминология, терминоведение, терминография..... 25	25
Лексический уровень системы языка	26
Термины и терминология в лексике языка	27
Терминография.....	29
Типология словарей	30
Современная языковая ситуация. Проблемы деловой прозы.....	31
Терминологические словари.....	33
Глава 4. Терминология и общеупотребительная лексика в переводе..... 34	34
Общеупотребительная и общенаучная лексика	34
Переводческая классификация общенаучных слов.....	37
РАЗДЕЛ II. СТАНОВЛЕНИЕ И РАЗВИТИЕ МАШИННОГО ПЕРЕВОДА..... 39	39
Глава 5. Появление идеи машинного (автоматического) перевода	39
Как начинался машинный перевод	39
Основные этапы развития машинного перевода.....	40

Глава 6. О теории и практике машинного перевода	45
Двенадцать проблем машинного перевода	45
«Технические проблемы», оценка качества и перспективы.....	48
Глава 7. Эффективность моделирования перевода и понятие «подъязык»	49
Естественный язык и «подъязыки».....	49
Системность в лексике	50
Функционально-коммуникативная модель микроподъязыка.....	52
Глава 8. Модели перевода и возможности их формальной реализации	64
Комбинаторная и количественная лингвистика	64
Лингвистические основы исследования	67
Методика количественного исследования.....	70
Моделирование с помощью распределений.....	71
Интерпретация статистических распределений.....	75
Модели «человеческого» («ручного») перевода.....	78
Модели машинного (автоматического) перевода	83
Глава 9. Моделирование в компьютерной лингвистике	86
Моделирование языковых сущностей и человеческого мышления.....	87
Искусственный интеллект	90
Представление знаний	94
Знание как объект моделирования	97
Лингвистический алгоритм, нечеткие множества, порождение речи.....	100
Лексико-грамматический анализ текста в формальном аспекте.....	109
 РАЗДЕЛ III. МАШИННЫЙ ПЕРЕВОД КАК ЦЕНТРАЛЬНАЯ ПРОБЛЕМА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА	113
Глава 10. Значение идеи машинного перевода	113
Машинный перевод и теория языка.....	113
К истории машинного перевода	117
Современное состояние машинного перевода	118
Основные проблемы машинного перевода.....	119
Глава 11. Машинный перевод как центральная проблема искусственного интеллекта	127
Место машинного перевода в ряду интеллектуальных задач.....	127
Машинный перевод как «побочное следствие» преобразований.....	131
Возможности такого решения.....	132
Системы машинного перевода АМПАР и СПРИНТ	133

Глава 12. Общая стратегия разработки систем машинного перевода на основе модели переводных соответствий.....	134
Общие принципы построения модели	134
Блок-схема алгоритма машинного перевода на основе переводных соответствий	135
Различие в подходах к моделированию перевода	137
Единица перевода	139
Концепция модели переводных соответствий для машинного перевода.....	147
Контекстное разрешение лексической многозначности в машинном переводе.....	158
Заключение.....	163
Список литературы.....	168

Учебное издание

Марчук Юрий Николаевич

Модели перевода

Учебное пособие

Редактор *Н. М. Тимакова*

Технический редактор *Е. Ф. Коржуева*

Компьютерная верстка: *А. В. Бобылёва*

Корректор *В. Н. Рейбекель*

Изд. № 101115027. Подписано в печать 17.08.2010. Формат 60×90/16.

Гарнитура «Ньютон». Печать офсетная. Бумага офс. № 1. Усл. печ. л. 11,0.

Тираж 1 000 экз. Заказ № 30660.

Издательский центр «Академия». www.academia-moscow.ru

125252, Москва, ул. Зорге, д. 15, корп. 1, пом. 26Б.

Адрес для корреспонденции: 129085, Москва, пр-т Мира, д. 101В, стр. 1, а/я 48.

Тел.: (495) 648-0507, факс: 616-0029.

Санитарно-эпидемиологическое заключение № 77.99.60.953.Д.007831.07.09 от 06.07.2009.

Отпечатано в соответствии с качеством предоставленных издательством
электронных носителей в ОАО «Саратовский полиграфкомбинат».

410004, г. Саратов, ул. Чернышевского, 59. www.sarpk.ru