

ОПЫТЫ МАШИННОГО ПЕРЕВОДА¹

ОСНОВНЫЕ ПРОБЛЕМЫ МАШИННОГО ПЕРЕВОДА

(вводная статья)

Перевод текстов с одних языков на другие имеет большое значение для развития культуры и является необходимым звеном в обмене культурными ценностями, созданными различными народами. На перевод текстов затрачивается огромный труд. Это ставит на очередь вопрос об автоматизации и механизации процесса перевода. В последнее время для механизации перевода открылись большие возможности, связанные с созданием автоматических электронных вычислительных машин.

Цель настоящей статьи — освещение возможностей математических машин с точки зрения использования их для перевода и формулировка ряда лингвистических вопросов, решение которых позволило бы повысить эффективность машинного перевода.

Первые опыты машинного перевода были сделаны Л. Достертом, П. Гарвином и П. Шериданом в 1954 г. в США². У нас в Советском Союзе опытный машинный перевод с английского на русский при помощи машины БЭСМ был осуществлен в конце 1955 г. И. К. Бельской, Л. Н. Королевым, И. С. Мухиным, Д. Ю. Пановым, С. Н. Ракумовским³.

Возможности вычислительных машин

Назначение вычислительных машин состоит в переработке поступающей в них информации по определенным правилам. Поступающая информация должна быть закодирована при помощи чисел и введена в «память» машины, т. е. каждый знак вводимой информации должен быть вполне однозначным способом изображен при помощи чисел. Для переработки информации служит так называемая программа, которая тоже хранится в «памяти» машины. Машина выполняет переработку информации в строгом соответствии с программой. Обычно исходной информацией являются числовые данные задачи, а программа реализует счет, необходимый для решения предложенной задачи, однако можно строить программы для реализации других видов переработки информации. Основное условие для того, чтобы машина могла переработать информацию некоторого заданного вида, состоит в том, чтобы эта переработка осуществлялась по четко описанным правилам. Эти правила должны быть связаны между собой логическими условиями так, чтобы в каждом отдельном случае было ясно, какое правило должно действовать. Каждое правило само по себе должно быть совершенно отчетливым, чтобы его можно было применять буквально.

Необходимо отметить очень большие технические возможности современных вычислительных машин. Например, эти машины выполняют несколько тысяч арифметических операций в одну секунду. «Память» машины устроена так, что она может хранить информацию очень большого объема — сотни тысяч знаков. Программа, перерабатывающая информацию, может быть самой разнообразной. Эти технические возможности позволяют использовать вычислительные машины не только по прямому назначению, т. е. для производства вычислений, но и для разных других видов переработки информации, в частности для перевода с одних языков на другие.

Для того чтобы заставить машину переводить текст с одного языка на другой, нужно: 1) выработать систему правил, употребление которых в сочетании со специ-

¹ С этого номера журнал начинает печатать серию статей, посвященных вопросам машинного перевода.

² См. В. П. Берков и Б. А. Ершов, О попытках машинного перевода, ВЯ, 1955, № 6.

³ См. Д. Ю. Панов, Автоматический перевод, М., 1956.

аллимым словарем позволит механически переводить текст; 2) записать эти правила, а также словарь на «языке машины», т. е. составить программу, реализующую эти правила. Программа и словарь кодируются специальным образом и вводятся в машину. После этого машина подготовлена к переводу.

Наиболее трудным из указанных вопросов является первый. Он требует непрерывной совместной работы математиков и лингвистов. Следует заметить, что машина действует согласно введенным в нее правилам чисто формально. Она не может апеллировать к смыслу переводимого текста и использует только его начертание.

Проблема машинного перевода представляет большой научный интерес, так как требует нового подхода к изучению языков и взаимоотношений между способами выражения одной и той же мысли на разных языках.

Против некоторых предрассудков

Для того чтобы идеи и методы машинного перевода и его обоснования как со стороны математика, физика и техники, так и со стороны лингвистики были понятны и продуктивны, необходимо прежде всего рассеять некоторые предрассудки.

1. Некоторые лингвисты считают неправомерным соединение перевода и машины. Данное возражение понятно, если исходить из различных психологическими и эстетическими школами считать язык творческим актом, искусством. Однако язык прежде всего — орудие общения, его основная функция — коммуникативная, именно поэтому вполне возможно представить его в удобной для машинного перевода форме.

Никто не отнимает у лингвистов права думать, что язык может быть и «первоэлементом литературы» (Горький) и что речевой акт может быть «творческим». Но тем не менее язык может просто сообщать, и то, что он сообщает, можно и нужно довести до адресата любым удобным способом, в том числе — при помощи машины.

2. Второе недоразумение происходит от того, что применительно к машинному переводу некоторые термины лингвистики употребляются в необычном значении, например, такие, как «основа», «окончание», «фразиды слов» и многие другие. В следующих далее конкретных статьях необходимости введения особых терминов разъясняется. Это вызывает технические условия работы машины и не меняет предмета лингвистики. Речь идет не о пересмотре общепринятой лингвистической терминологии, а о том, как проще «задать» машине языковую, объективно выделенную данность. Против подобного переименования трудно выдвинуть сколько угодно серьезные возражения, если оно логически оправдано и удобно для машины.

Возьмем, к примеру, рассматриваемое ниже (в статье О. С. Кулагиной и И. А. Мельчука), в правилах перевода с французского языка на русский, чисто графическое в ряде случаев выделение основы и окончания. Конечно, с таким чисто графическим выделением морфем мы имеем дело лишь в языке, который пользуется буквенным письмом. Иначе обстоит дело при оперировании, например, с китайским языком, который пользуется иероглифическим письмом, где каждый знак передает или слово, или часть его, соответствующую по значению нашей морфеме. Впрочем работа по машинному переводу применительно к китайскому языку в настоящее время только начата, и останавливаться на ней мы пока не будем.

Поскольку, как мы говорили выше, при машинном переводе язык берется в его коммуникативной функции и поскольку отношения между двумя языками должны быть определены чисто формально¹, правомерен и тот подход к языку, который зачастую неодобрительно называют «формализмом» или «структурализмом».

Вопрос о том, что может сообщить язык, тесно связан с выражаемым и выражающим, обозначающим и обозначающим. В этом отношении язык можно и должно рассматривать как систему информации. Для машинного перевода важно, как структурно-формально дать элементы переводимого, чтобы через машину получить адекватные элементы переведенного.

Для решения этих задач прежде всего надо исключить психологию и физиологию, которые исследуют акт речи как психо-физиологический процесс, так как то, что интересует психологов и физиологов, для установления правил кибернетического перевода не нужно. Наоборот, в содружестве математиков и лингвистов необходимо включить логику, так как здесь вопрос о данностях и их структурных характеристиках тесно связан с вопросом об отношениях. Для машинного перевода в равной степени необходимы и решение этих двух вопросов, и их обязательная связь. Отсюда следует, что необходимы: 1) точная регистрация всех фактов языка, с которого и на который переводит, и их качественных и исчислимых определяемых данных; 2) определение всех возможных отношений и сочетаний, в которые эти факты вступают;

¹ Машина имеет дело лишь с определенной последовательностью языковых единиц, выраженной графическими средствами данного языка или последовательностью звуков речи. В настоящее время ставится проблема перевода звучащей речи одного языка на звучащую речь другого языка, однако в данной статье последняя проблема не рассматривается.

3) формулировка «правил», которые можно задать машине, чтобы она скорейшим образом, учитывая количество «ошибочных» ответов, смогла дать удовлетворительный, «правильный» ответ.

Работа по решению задачи формализации отношений между языками должна опираться на те лингвистические теории, которые в исследовании языка исходят из его формальной стороны. В современном языкознании это — различные направления структурализма.

Изучение языка с этой точки зрения мы находим уже в так называемом формализме, система которого в наиболее четком виде была впервые разработана Ф. Ф. Фортунатовым и его учениками. Вряд ли случайно то обстоятельство, что при подготовке правил перевода с английского языка на русский наиболее выгодной оказалась последовательная разбивка текста на двучленные группы, в каждой из которых один член грамматически зависит от другого. Ведь эта разбивка представляет собой по существу разбивку текста на словосочетания (в понимании Ф. Ф. Фортунатова).

Некоторые общелингвистические вопросы машинного перевода

Разработка правил машинного перевода с одного языка на другой ставит перед лингвистами ряд интересных проблем, которые раньше или не возникали вообще, или ставились в ином плане. Эти проблемы относятся как к переводу с различных языков и на различные языки, так и к принципам анализа текста на некотором данном языке. Уже сейчас можно сказать, что различие в типе грамматического (и именно морфологического) строя языка обуславливает некоторые различия в характере переводческих правил для данного языка (при сохранении, конечно, некоторых общих принципов построения правил). Необходимость таких различий была обнаружена уже при разработке правил перевода с французского на русский и с английского на русский, поскольку в английском языке отсутствует ряд форм, имеющих во французском языке, вследствие чего гораздо уже представлено согласование как форма связи между членами словосочетания.

Лингвистической задачей машинного перевода является выработка такой системы правил, которая позволила бы, имея некоторый текст на данном языке, однозначно определить, какой текст другого языка должен быть адекватен по смыслу данному тексту. Любой язык имеет словарный состав и грамматический строй. Поэтому любой текст любого языка содержит слова, связанные между собой по нормам грамматического строя данного языка. Различные грамматические связи служат для выражения различных отношений между понятиями, выражаемыми словами. Поэтому правила перевода с одного языка на другой должны содержать как правила перевода слов, так и правила передачи грамматических средств одного языка, выражающих некоторые отношения, грамматическими средствами другого языка, выражающими те же отношения.

Ввиду различий, существующих между языками, невозможно построить некую единую универсальную систему правил перевода с любого языка на любой, но каждая система правил, какая может быть построена, должна связывать лишь два конкретных языка, притом лишь в одном направлении. Так, например, может быть построена одна система правил для перевода с французского языка на русский, другая — с русского на французский, третья — с английского на русский и т. д. Но в то же время могут быть определены и некоторые общие принципы, т. е. некоторые общие правила для построения различных конкретных систем правил, точно так же, как при всем различии грамматической структуры разных конкретных языков земного шара существуют некоторые общие принципы изучения грамматической структуры языка, на которые, в частности, опирается и классификация языков по их грамматическому типу.

При наличии большого количества языков, которые должны быть связаны переводческими отношениями, встает вопрос, как более целесообразно строить правила перевода для двусторонней связи каждого данного языка с другим в пределах определенной группы. Возможно, что целесообразно сначала переводить на один какой-нибудь язык, через посредство которого осуществлять перевод на другие языки. Операция перевода в этом случае, повидимому, несколько замедляется, так как любой перевод осуществляется не непосредственно, а двумя этапами: с первого конкретного языка на язык-посредник, а затем с языка-посредника на другой конкретный язык, но зато требуется меньшее количество систем правил, чем при переводе непосредственном. В качестве языка-посредника можно избрать один из конкретных языков, но возможно построение и некоторого специального языка. Этот специальный язык может быть языком определенных логических символов, позволяющих передать точно содержание переводимого текста, или же логическим языком, в основе которого лежат слова и формы реального языка, отобранные таким образом, что он будет характеризоваться теми же свойствами, что и логический язык символов. Возникает вопрос, не представляет ли подобный язык известных выгод в любых условиях перевода, независимо от количества языков, находящихся в переводческих отношениях, по-

сколько возможно, что легче точно определить формальные соответствия между таким специальным языком и любым конкретным, чем между двумя конкретными. Этот вопрос нуждается в специальном исследовании.

Правила перевода с любого данного языка на любой другой содержат, как уже было сказано, правила перевода слов и правила перевода грамматических средств (именно средств, а не форм, так как грамматической формой принято называть форму одного отдельного слова, каковой в некоторых языках может и не быть). Переводы слов содержатся в закодированном в порядке алфавита данного языка словаре. Но должны существовать и некоторые правила перевода слов, так как лишь сравнительно редко одно слово данного языка соответствует одному и только одному слову другого языка. Большей частью слово данного языка соответствует нескольким словам другого языка (в этом легко убедиться, взглянув в любой двуязычный словарь), и должен быть однозначно определен выбор слова второго языка из нескольких возможных.

Кроме того, в любом языке имеются особые фразеологические сочетания, значение компонентов которых не тождественно их значению вне этих сочетаний. В правилах для перевода с французского языка на русский, разработанных О. С. Кулагиной и И. А. Мельчуком, предусмотрены и такие случаи, когда французское слово может быть переведено одновременно двумя русскими словами, из которых читатель, знакомый с той специальной областью, к которой относится переводимый текст, сам выберет нужное по смыслу. Но такой двойкий перевод допускается лишь на настоящей стадии работы как временная мера; в дальнейшем будут определены правила, согласно которым выбор одного из двух значений должна осуществлять машина.

Правила передачи грамматических средств должны учитывать как морфологическую, так и синтаксическую сторону обоих языков. В тексте языка, с которого переводят, у каждого слова должны быть вычленены морфемы словоизменения, если они есть в языке. Это необходимо, во-первых, для поиска слова в словаре, где, совершенно очевидно, слово излишне давать во всех его формах и где приводятся лишь основы всех изменяемых слов; во-вторых, для определения (по этой словоизменяющей морфеме) синтаксического отношения данного слова как зависимого в словосочетании к другому слову того же словосочетания или же для определения значения глагольной формы (времени, наклонения, залога, иногда лица). При выделении основы должен быть учтен ее звуковой (точнее фонемный) вид (переданный на письме определенной графической формой), так как в ряде языков имеются так называемые чередования. Это необходимо, во-первых, для поиска в словаре, где любое слово дается в одном, принятом как основном, виде основы (с указанием на наличие других видов основы), во-вторых, для определения синтаксического отношения данного слова к другому, включая предикативные формы глагола (см. выше), так как соответствующие значения в некоторых языках выражаются и чередованиями. В тексте языка, на который переводят, слова, найденные в словаре также в виде основ (если они не входят в состав устойчивых фразеологических сочетаний), должны быть снабжены словоизменяющими морфемами, выражающими грамматические значения, свойственные словам языка, с которого переводят (если такие морфемы есть в языке, на который переводят). Основам же нужно придать вид, свойственный формам слов, соответствующих по значению отношениям, выраженным в тексте языка, с которого переводят (если в языке, на который переводят, имеют место изменения основы). Синтаксическая сторона должна быть принята во внимание в двойном отношении. Во-первых, один и те же отношения между членами словосочетания в языке, с которого переводят, и в языке, на который переводят, могут быть выражены в одном морфологическими, а в другом синтаксическими средствами (т. е. посредством так называемых служебных слов или посредством порядка следования членов словосочетания — употребляем последний термин в значении, приданном ему Фортунатовым), и при переводе должны быть использованы именно средства, свойственные языку, на который переводят. Во-вторых, при переводе в ряде случаев должен быть изменен порядок слов согласно нормам того языка, на который переводят. Порядок слов, как известно из только что сказанного, имеет особо существенное значение в тех языках, где он грамматикализован, т. е. где ряд отношений в пределах словосочетания выражается исключительно посредством порядка слов. Но порядок слов должен быть принят во внимание и в ряде случаев изменен сравнительно с текстом языка, с которого переводят, и при переводе на такой язык, в котором порядок слов не грамматикализован (как, например, в русском). Слова должны быть расположены в привычном для данного языка порядке. При этом следует заметить, что в языке, где порядок слов не грамматикализован, последний может быть использован для выделения соответствующих членов, в то время как в языках с грамматикализованным порядком это выделение может осуществляться другими средствами, например путем сопровождения выделяемых членов различными частицами. При переводе эти частицы могут опускаться, а выделение соответствующего члена может осуществляться посредством порядка слов (так, в русском языке выделение члена осуществляется преимущественно вынесением его на конец предложения, в особенности в сочетании с дистантным положением членов словосочетания, в которое данный член входит).

Особенности письменного машинного перевода

При анализе текста применительно к некоторому данному языку прежде всего возникает вопрос о том, с каким отрезком мы имеем дело и на какие единицы он последовательно разбивается. Существенно определить, какой отрезок переводимого текста содержит грамматические связи, осуществляющиеся лишь внутри него и не переходящие за его границы. Определение границ такого отрезка не столь просто. Известную помощь оказывает пунктуация, но не во всех случаях. Практически берется отрезок, содержащийся между двумя точками, т. е. тот, который синтаксически представляет собой предложение, возможно, так называемое сложное предложение. Большая часть грамматических связей действительно заключается внутри него. Но, во-первых, внутри этого отрезка могут находиться различные отрезки, представляющие собой некоторые единицы, никаких грамматических связей между которыми нет. Это относится, прежде всего, к последовательности предложений, не связанных между собой никакими грамматическими средствами (союзами, относительными словами) и лишь разделенных на письме запятыми. Во-вторых, некоторые связи выходят все же за пределы отрезка, графически ограниченного точками. Это относится к местоимению типа *он* (*она*, *оно*), выбор форм которого в ряде языков может быть обусловлен принадлежностью к определенной категории (рода, лица и т. п.) существительного, лежащего за пределами данного отрезка.

Что касается единиц, расположенных внутри отрезка, ограниченного точками, то их выделение в некоторых случаях также осуществляется при помощи пунктуации, иногда же пунктуация оказать помощи не может, что в определенных случаях зависит от принятой для данного языка пунктуационной системы. Так, например, из соответствующих друг другу по значению французского оборота с относительным местоимением *qui* и русского с *который* русский оборот выделяется запятыми, выступающими как внешние показатели его границ. Французский оборот запятыми не выделяется, начальная его граница определяется формой *qui*, конечная же должна быть определена другими способами (не при помощи пунктуации и не при помощи особого слова, подобно тому как определяется начальная граница).

Нижеследующие статьи посвящены описанию некоторых специальных вопросов машинного перевода и работ по его реализации.

П. С. Кузнецов, А. А. Ляпунов, А. А. Реформатский

МАШИННЫЙ ПЕРЕВОД С ФРАНЦУЗСКОГО ЯЗЫКА НА РУССКИЙ

В настоящей работе излагаются принципы построения словаря и «грамматики» (свода специальных правил) для машинного перевода с французского языка на русский научного текста (из области математики), а также описывается сам процесс перевода.

Для перевода необходимо иметь следующие данные: 1) «Словарь основ»; 2) «Словарь оборотов»; 3) «Таблицы перевода предлогов»; 4) «Правила различения омонимов»; 5) «Таблицы окончаний» (французских и русских); 6) «Группа анализирующих правил»; 7) «Группа синтезирующих правил». По мере надобности эти данные вводятся в «память» машины.

I

Для машинного перевода математических текстов с французского языка на русский нами был подготовлен особый словарь на основе анализа словарного состава различных математических текстов. Подсчитывались как слова, так и словосочетания. Было рассмотрено 20 500 слов, среди них оказалось 2300 различных. Из этих слов были отобраны те, которые встретились не менее четырех раз; их оказалось около 1000. К этой тысяче слов были добавлены без статистических подсчетов термины, необходимость которых была очевидна (их около 50), а также все служебные слова. В результате получился словарь приблизительно на 1100 слов.

Особенность этого словаря в том, что в нем помещены не слова, а основы. Под термином «основа» здесь понимается часть слова, которая остается графически неизменной во всех его формах, т. е. не совсем то, что обозначает этот термин в обычной грамматике. Часть слова, оставшуюся после отделения «основы», мы называем «окончанием». Так, для слова *travail* в нашем словаре основой будет *trava-* [так как у этого слова две формы: *trava-il* (ед. число) и *trava-ux* (мн. число)], для *parler* основа *parl-*, для *finir* основа *fini-* и т. д. Отметим, что для большинства французских существительных «основа» в нашем понимании совпадает с формой единственного числа.