

ОПЫТЫ МАШИННОГО ПЕРЕВОДА

НЕКОТОРЫЕ ВОПРОСЫ СИНТАКСИСА В СВЯЗИ С МАШИНЫМ ПЕРЕВОДОМ
С АНГЛИЙСКОГО ЯЗЫКА НА РУССКИЙ

Настоящая статья представляет собой описание некоторых предварительных исследований, которые ведутся для выработки правил машинного перевода с английского языка на русский. Надо отметить, что основные трудности проблемы машинного перевода связаны не с технической, а с лингвистической его стороной. Лингвистика должна предложить такое решение этой проблемы, которое обеспечивало бы возможность полного перевода разнообразного материала без предварительной его обработки и перестройки в зависимости от лексических и грамматических особенностей того языка, на который делается перевод.

Лингвистическая проблема включает в себя следующие вопросы: 1) орфографию (этот вопрос важен в связи с тем, что машинный перевод, по крайней мере на первых порах, имеет дело с графической формой языка); 2) морфологию; 3) лексикологию (составление словарей особого типа); 4) фразологию (в этом разделе важнее всего вопросы полисемии и идиоматики); 5) синтаксис.

При переводе с английского языка наибольшие трудности представляют два последних вопроса, так как они касаются тех сторон и явлений языка, которые создают самые серьезные препятствия для проведения формального анализа. Особенно большие затруднения возникают при разрешении синтаксических проблем. Это объясняется тем, что для формально-морфологического анализа имеется очень мало данных: английский язык беден словосменительными формами, среди грамматических суффиксов наблюдается омонимия, очень продуктивно такое средство словообразования, как конверсия.

Цель проведенного нами исследования состояла в следующем: выяснить, какие модели простейших свободных сочетаний слов существуют в английском и русском языках, а затем установить соответствие между ними. Если удастся установить тип русских словосочетаний, соответствующих каждому английскому типу, то перевод можно будет осуществить следующим образом: английское предложение подвергается анализу (выделение простейших свободных сочетаний), а затем из русских сочетаний, соответствующих данным английским, синтезируется русское предложение. При этом мы допускаем, что любое предложение может быть разложено на простейшие словосочетания. Это допущение подкрепляется синтаксической теорией так называемого синтагматического анализа¹, согласно которой основной синтаксической единицей является синтагма. Такой анализ дает возможность разложить английское предложение на синтагмы, а затем синтезировать из отдельных переведенных синтагм русское предложение.

Но отношение определяемого и определяющего, характеризующее синтагму, не исчерпывает всех отношений между словами в предложении. Существуют, например, сочинительные сочетания (*белый и красный; боится, но идет*). При анализе и синтезе предложений необходимо учитывать такие сочетания, также как и подчинительную связь, выходящую за пределы одного предложения (например, абсолютный причастный оборот в английском языке). Поэтому мы рассматриваем не только синтагмы, но все возможные типы свободных сочетаний слов в английском и русском языках.

Что представляют собой эти типы, или модели, свободных сочетаний слов? Это не группы слов как таковых, а сочетания, состоящие из классов слов, расположенных в определенном порядке и имеющих определенное грамматическое оформление. Например, прилагательное + существительное, согласованные в роде, числе и падеже (*глубокий снег*); переходный глагол + существительное в вин. падеже (*писать письмо*). Необходимо определять, какие классы слов могут соединяться между собой, в какой последовательности и при помощи каких грамматических средств. Чтобы выяснить эти

¹ См.: С. И. Карпневский, Повторительный курс русского языка, М.—Л., 1928, стр. 24—36 (гл. II); А. А. Реформатский, Введение в языковедение, М., 1955, стр. 252—259.

вопросы, прежде всего нужно распределить все слова по классам. Традиционная классификация по частям речи оказывается неудовлетворительной, так как в ряде случаев она опирается на значение слова, а не на чисто формальные признаки. (Например, в такие части речи, как местоимения и наречия, объединяются формально разнородные группы слов). Поэтому для целей машинного перевода мы выбираем другую классификацию: слова разбиваются на классы по признаку одинакового окружения¹.

Выделенные классы слов частично совпадают с теми, которые рассматриваются в так называемом структуралистическом языкознании, в частности, с формальными классами Блумфилда и Фриеса². В один класс объединяются все слова, занимающие ряд одинаковых положений в предложении. Например, к классу глаголов в английском языке принадлежат слова, характеризующиеся следующими особенностями: им может предшествовать существительное или личное местоимение в общем падеже, с которым они согласуются в числе; за ними может следовать существительное или личное местоимение без предлога (последнее в объектном падеже), существительное или личное местоимение с предлогом, предлогообразное наречие, прилагательное, причастие I на *-ing* и т. д.

Всего было выделено 19 классов слов в английском языке и 17 — в русском. При этом внутри некоторых классов возможны подклассы. Условимся обозначать каждый класс определенным индексом. Тогда наша классификация слов будет выглядеть следующим образом:

- | | |
|---|---|
| 1. 1 — существительные | к) 2 ^{ing} — причастие I (активное) |
| 2. P — личные местоимения | л) 2 ^{ed} — причастие II (пассивное) |
| 3. 1 ^{so} — безличное местоимение <i>it</i> | м) 2 ^m — модальные глаголы |
| 4. 1 th — обстоятельственные наречия или существительные, употребляемые в функции обстоятельств времени (например, <i>Next morning we shall go to the country</i>). | 6. 3 — прилагательные (в русском языке выделяется подкласс кратких прилагательных — 3 ^e) |
| 5. 2 — глаголы | 7. 4 — обстоятельственные наречия |
| а) 2 ^a — глаголы, после которых употребляется так называемый объектный предикативный член | 8. 4 ^x — качественные наречия, большей частью на <i>-ly</i> (в английском) и <i>-o</i> (в русском) |
| б) 2 ⁺ — переходные глаголы | 9. 4 ^w — предлогообразные наречия (только в английском языке) |
| в) 2 ⁻ — непереходные глаголы | 10. Δ — наречия меры и степени |
| г) 2 ^b — глагол <i>to be</i> (соответственно русский «быть») | 11. N ^c — порядковые числительные |
| д) 2 ^a — глагол <i>to do</i> (только в английском языке) | 12. N — притяжательные и указательные местоимения |
| е) 2 ^x — глагол <i>to have</i> (только в английском языке) | 13. N ₁ — неопределенные местоимения <i>much, many, few, any, some, little</i> , отрицательное местоимение <i>no</i> в английском языке; <i>много, мало, несколько, ничего</i> в русском и количественные числительные |
| ж) 2 ^l — глаголы <i>shall, will</i> (только в английском языке) | 14. D — артикли (в английском языке) |
| з) 2 ^w — глаголы <i>should, would</i> (только в английском языке) | 15. F — предлоги |
| и) 2 ^{ll} — глагол <i>let</i> | 16. E — сочинительные союзы |
| | 17. J — подчинительные союзы |
| | 18. I — все вопросительные слова |
| | 19. C — отрицательная частица <i>not</i> в английском языке и <i>не</i> — в русском. |

Исходным материалом для выделения элементарных свободных сочетаний слов послужили типичные по строению английские предложения, собранные Фриесом³ и перечисленные в его книге, а также образцы синтаксических построений, приведен-

¹ Занимаясь анализом языка для осуществления механического перевода, не всегда приходится придерживаться положений научной лингвистики. Машинный перевод — эта прикладная область языкознания — основывается главным образом на соображениях выгоды и удобства обработки языкового материала машиной. Поэтому, когда положения научной лингвистики не противоречат удобству анализа, производимого при машинном переводе, мы всецело опираемся на них. Но если произвольный подход к языковому материалу лучше служит нашим целям, мы выбираем его. Отсюда — довольно частые расхождения с традиционными понятиями лингвистики.

² См.: L. Bloomfield, *Language*, New York, 1948; Ch. C. Fries, *The structure of English*, New York, 1952.

³ См. Ch. C. Fries, указ. соч.

ные в книге О. Есперсена «Аналитический синтаксис»¹. Эти английские синтаксические типы были, по возможности стандартно, переведены на русский язык. Русские предложения также были проанализированы, а затем было произведено сопоставление английских и русских словосочетаний.

Выделились в основном двучленные словосочетания, т. е. самые простые. Но если два элемента английского предложения не давали возможности установить соответствие с русским языком — в сочетании объединялось большее количество слов. Это наблюдается в следующих случаях:

1. В абсолютном причастном обороте:

англ. русск.

$12^{1st} 12 = J 1212$

The rain having ruined my hat, I had to buy a new one = Так как дождь испортил мою шляпу, я должен был купить новую.

2. При использовании объектного предикативного члена:

$$2^* 12^{\pm} = 2^* J 12^{\pm}$$

saw the boy run = видел, как мальчик бежал.

3. В сочетаниях с безличным *it*:

$$1^{so} 2^b 3 = 2^{(b)} 3^{e2}$$

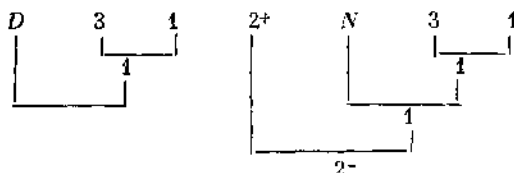
It was cold = было холодно.

Как известно, в каждом словосочетании имеется главный компонент — стержневое, грамматически господствующее слово, которое подчиняет себе второе слово. Поэтому все сочетание можно упрощать, заменяя его главным компонентом, подобно тому как это делается при синтагматическом анализе. Например:

$$31 \rightarrow 1, 1F1 \rightarrow 1, 24^x \rightarrow 2, \Delta 3 \rightarrow 3.$$

При помощи такого приема достигается еще большая степень обобщения при анализе и создается возможность последовательно устанавливать связи между всеми словами в предложении, не только между теми, которые находятся в контактном положении, но и теми, которые не стоят рядом друг с другом, хотя реально связаны между собой.

Например: *An old man takes his shabby coat*.



Сочетание двух слов *shabby* и *coat* (3 1) упрощается в один комплекс *shabby coat* (1); это новое образ. ание (1) объединяется с *his* (N) и дает еще более сложный субстантивный комплекс *his shabby coat* (1), который в свою очередь объединяется с глаголом *takes* (2⁺) и подчиняется ему, вливаясь в его формулу (2⁺ 1 - 2⁻). Подобным же образом постепенно упрощается последовательность слов *an old man*, образуя в конечном итоге субстантивный комплекс (1). Общий результат анализа — 1 2⁻. Дальнейшее объединение частей этого словосочетания не является необходимым, так как оно уже не выясняет, между какими словами существуют синтаксические связи. Такой метод упрощения словосочетаний оказывается очень полезным. Он позволяет производить «свертывание» английского предложения и затем «развертывание» русского³.

Большинство выделенных английских словосочетаний однозначны, т. е. имеют по одному русскому соответствию каждое. Но иногда встречаются и многозначные сочетания. Например:

англ. русск.
 $2^{\pm} 3 = 2^{\pm} 3$ или $2^{\pm} 4^x$

¹ См. О. Jespersen, *Analytic syntax*, Copenhagen, 1937.

² Скобки указывают на то, что связка *быть* употребляется не всегда: в настоящем времени она опускается.

³ При этом слова классов 1th и 4 не учитываются в сочетаниях и формулах, так как их синтаксическая связь с отдельными словами в предложении ослаблена. Они относятся ко всему предложению в целом.

He looks pale = Он выглядит бледным

The stone lay deep in the water = Камень лежал глубоко в воде

$1F1 = 11$ (род. падеж) или $1F1$

A group of children = группа детей

A book with pictures = книга с картинками.

Очевидно, для этих случаев полисемии или омонимии нужно составить дополнительные правила различения. Например, при синтезе русского предложения выбор формулы $1F1$ или 11 (род. падеж) зависит от того, является данный предлог предложением *of* или нет. В случае предлога *of* выбираем формулу 11 (род. падеж), в случае другого предлога — $1F1$.

При анализе английского предложения возникают трудности, связанные с тем, что иногда одну и ту же группу слов можно разбить на словосочетания по-разному. Например:

The old man
D 3 1

Индексы этой последовательности можно сгруппировать двумя способами:

1. $D \quad 3 \quad 1$ (Формула $D3-1$ введена для случаев типа

$\begin{array}{|c|} \hline 1 \\ \hline \end{array}$

The wounded were counted, где прилагательное

$D \quad 3 \quad 2^b \quad 2^+ \quad ed$
 $\begin{array}{|c|} \hline 1 \\ \hline \end{array} \quad \begin{array}{|c|} \hline 2 \\ \hline \end{array}$

должно переводиться на русский язык во множественном числе).

2. $D \quad 3 \quad 1$
 $\begin{array}{|c|} \hline 1 \\ \hline \end{array}$
 $\begin{array}{|c|} \hline 1 \\ \hline \end{array}$

Второе членение является единственно верным. Чтобы избежать неправильного членения, нужно, чтобы преобразование $31 - 1$ производилось раньше, чем $1:3-1$.

Следовательно, должна соблюдаться известная очередность применения формул. Поэтому все словосочетания были разделены на несколько групп, и группы расположены в определенной, строгой последовательности. Таким образом, анализ английского предложения будет производиться в порядке следования групп словосочетаний: если в анализируемом предложении не обнаружено словосочетаний, отнесенных к первой группе, переходим ко второй группе, третьей и т. д. Нарушать последовательность групп, например, возвращаться к первой после третьей, нельзя.

Примеры распределения формул по группам:

I $\begin{cases} 4^x 4^x \rightarrow 4^x \\ \Delta 4^x \rightarrow 4^x \\ 4^x \Delta \rightarrow 4^x \end{cases}$

II $\begin{cases} 4^x 3 \rightarrow 3 \\ C 3 \rightarrow 3 \\ N_1 3 \rightarrow 3 \\ 3 3 \rightarrow 3 \\ \Delta 3 \rightarrow 3 \end{cases}$

III $\begin{cases} 3 1 \rightarrow 1 \\ 2^{-ing} 1 \rightarrow 1 \\ 2^{\pm ed} 1 \rightarrow 1 \\ N^c 1 \rightarrow 1 \\ N 1 \rightarrow 1 \\ C 1 \rightarrow 1 \end{cases}$

IV $D 1 \rightarrow 1$

V $\begin{cases} 2^{+ing} 1 \rightarrow 2^{-ing} \\ 2^{+ing} P \rightarrow 2^{-ing} \end{cases}$

VI $\begin{cases} 2^* 12^{\pm} \rightarrow 2^{\pm} \\ 2^* P 2^{\pm} \rightarrow 2^{\pm} \\ 2^* 1 2^{\pm ing} \rightarrow 2^{\pm} \\ 2^* P 2^{\pm ing} \rightarrow 2^{\pm} \\ 2^* 1 2^{\pm ed} \rightarrow 2^{\pm} \\ 2^* P 2^{\pm ed} \rightarrow 2^{\pm} \end{cases}$

VII $1F1 \rightarrow 1$

VIII $\begin{cases} 2^b 1 \rightarrow 2^- \\ 2^b 3 \rightarrow 2^- \\ 2^b 4 \rightarrow 2^- \end{cases}$

IX $\begin{cases} 2^+ 1 \rightarrow 2^- \\ 2^+ P \rightarrow 2^- \\ 2^{\pm} 3 \rightarrow 2^- \end{cases}$

X $\begin{cases} 2^- F 1 \rightarrow 2^- \\ 2^- F P \rightarrow 2^- \\ 2^- F 2^{+ing} \rightarrow 2^- \end{cases}$

XI $\begin{cases} 2^+ F 2^{\pm} \rightarrow 2^{\pm} \\ 2^- F 2^{\pm} \rightarrow 2^{\pm} \end{cases}$

Как же осуществляется анализ английского предложения по типам словосочетаний, которые выделяются и преобразовываются в определенной последовательности? Каждому слову предложения приписывается индекс в соответствии с тем, к какому классу слов оно принадлежит. Индексы располагаются в той же последовательности, что и соответствующие им слова. Затем выясняется, есть ли в данной последовательности индексов сочетания, перечисленные в формулах первой группы. Если в этой последова-

тельности не обнаружено таких сочетаний, то переходит ко второй группе и т. д. до конца. Если в этой последовательности обнаружено одно или несколько сочетаний, перечисленных в формулах первой группы, то они упрощаются по соответствующим формулам. Индекс, полученный в результате упрощения данного сочетания, ставится на его место. Затем переходят ко второй группе и проделывают те же операции вплоть до последней группы. Например:

He found the bird flown
P 2* D 1 2^{-ed}

Номер группы	Формула	Анализируемая последовательность индексов
IV	$D1 \rightarrow 1$	$P 2^* D 12^{-ed} = P 2^* 12^{-ed}$
VI	$2^* 12^{-ed} \rightarrow 2^-$	$P 2^* 12^{-ed} = P 2^-$

После того как произведены преобразования всех групп, анализ английского предложения закончен. Следующий этап перевода — синтез русского предложения. Возвращаются к последнему из произведенных при анализе упрощений. Находит русскую формулу, соответствующую применявшейся английской. Производят указанное русской формулой преобразование и результат по-прежнему в полученную при анализе последовательность индексов. Затем переходят к следующей от конца группе и производят те же операции, рассматривая группы в обратном порядке. При этом русские словосочетания содержат в себе указания на грамматические формы входящих в них слов. Синтезируем, например, русское предложение, переводящее английское *He found the bird flown*.

Номер группы	Формула анализа	Формула синтеза	Синтезируемая последовательность индексов
IV	$D1 \rightarrow 1$	$\tilde{2}^{-1} \rightarrow \tilde{2}^{-} - J \tilde{1}^2$ (им.п.) $\tilde{2}^{\check{}}$	$\check{P}$ (им.п.) $\check{2}^{-} = \check{P}$ (им.п.) $\check{2}^{-} J \check{1}$ (им.п.) $\check{2}^{-}$
VI	$2^* 12^{-ed} \rightarrow 2^-$	$\tilde{1} \rightarrow \tilde{1}$	$\check{P}$ (им.п.) $\check{2}^{-} J \check{1}$ (им.п.) $\check{2}^{-} =$ $= \check{P}$ (им.п.) $\check{2}^{-} J \check{1}$ (им.п.) $\check{2}^{-}$

$\check{P}$ (им. п.) $\check{2}^{-} J \check{1}$ (им. п.) $\check{2}^{-}$ — схематическое изображение русского предложения, соответствующего данному английскому. Поставив на место индексов слова, которым эти индексы соответствуют, мы получим само предложение.

Такой анализ и синтез предложения по словосочетаниям можно производить, если о каждом слове предложения точно известно, к какому классу слов оно относится. В английском языке из-за распространенности лексико-грамматической омонимии часто нельзя однозначно установить принадлежность слова к определенному классу. Например, словоформы *work, works, sleep, sleeps, cry, cries* и т. д. могут являться и глагольными, и субстантивными словоформами. Этот факт языка должен быть отражен в словаре: следует снабдить каждое слово, имеющее лексико-грамматические омонимы, соответствующими указаниями. Так, слово *work* должно получить два индекса — 1 и 2⁻, *sleep* — 1 и 2⁻, *it* — 1³⁰ и *P, her* — *N* и *P* и т. д.

Следовательно, для осуществления анализа по словосочетаниям нужно предварительно выбрать из нескольких индексов данного слова один, подходящий в данном случае. Очевидно, подойдет тот индекс, который составляет с соседними индексами возможные в языке словосочетания, зафиксированные нами. Поэтому для того, чтобы выбрать индекс, анализируем окружение слова. Например, если нужно выбрать между 1 и 2[±], мы проверяем, не стоит ли перед данным словом *D, 3, N, N^e* и пр. Если стоит один из перечисленных индексов, то выбираем 1. Если таких индексов

¹ Знак ~ обозначает тождественность грамматической формы.

² √ и ∨ — согласование в лице, числе и роде.

нет, то проверяем не стоят ли перед данным словом 2^b , $2^w, 2^m, 2^a$, 4^x и пр. Наличие одного из перечисленных индексов показывает, что нужно выбрать $2^\pm$.

Иногда все индексы данного слова составляют вместе с индексами, стоящими вправо и влево от него, возможные словосочетания. В этом случае необходимы дополнительные правила различения. Так, для двойных индексов $1, 2^\pm$ применяют следующее дополнительное правило: если вправо и влево от данного слова, которому соответствуют индексы $1, 2^\pm$, до точки, запятой, подчинительного союза (J) или вопросительного слова (J) нет другого глагола ($2^\pm$), то выбираем $2^\pm$, если есть, то выбираем 1.

Вопрос о выборе индекса, т. е. об уничтожении в каждом данном случае лексико-грамматической омонимии, очень сложен. Он нуждается в доработке и уточнении. В этой связи известный интерес представляет метод определения принадлежности слова к тому или иному классу в данном контексте, предложенный Бар-Хиллелом¹. Возможны и другие вспомогательные приемы выбора правильного индекса слова.

Все вышеизложенное не претендует на исчерпывающую полноту и завершенность. Это пока рабочая гипотеза, которую предполагается проверить на большом материале при помощи машины.

Т. Н. Молошина

¹ См. Y. Bar-Hillel, A quasi-arithmetical notation for syntactic description, «Languages», vol. 29, № 1, 1953.