АКАДЕМИЯ НАУК СССР
ИНСТИТУТ СЛАВЯНОВЕДЕНИЯ

На правах рукописи

Н. А. ПАЩЕНКО

СИНТАКСИЧЕСКИЙ АНАЛИЗ И СОПОСТАВЛЕНИЕ СИНТАКСИЧЕСКИХ СТРУКТУР ЧЕШСКОГО И РУССКОГО ЯЗЫКОВ

(ПРИМЕНИТЕЛЬНО К МАШИННОМУ ПЕРЕВОДУ)

Автореферат диссертации на соискание ученой степени кандидата филологических наук

МОСКВА 1965

АКАДЕМИЯ НАУК СССР
ИНСТИТУТ СЛАВЯНОВЕДЕНИЯ

На правах рукописи

Н. А. ПАЩЕНКО

СИНТАКСИЧЕСКИЙ АНАЛИЗ И СОПОСТАВЛЕНИЕ СИНТАКСИЧЕСКИХ СТРУКТУР ЧЕШСКОГО И РУССКОГО ЯЗЫКОВ (ПРИМЕНИТЕЛЬНО К МАШИННОМУ ПЕРЕВОДУ)

Автореферат диссертации на соискание ученой степени кандидата филологических наук

Научный руководитель
кандидат филологических наук
В. В. ИВАНОВ

МОСКВА 1965

Результаты исследований в области машинного перевода, полученные в разных странах мира, показывают, что основные трудности решения этой важнейшей проблемы заключаются не в технике составления алгоритмов перевода и не в их машинной реализации, а в том, что при существующем уровне наших знаний о языке невозможно построение такого алгоритма, который обеспечивал бы получение адекватного и высококачественного автоматического перевода любого текста. Это объясняется, прежде всего, недостаточной изученностью естественных языков. Именно поэтому всё больше исследователей в этой области направляют сейчас свои усилия на изучение грамматической и семантической структур отдельных языков, на тщательное исследование и сопоставление фактов этих языков, а также на решение целого ряда задач лексикографического характера, связанных с составлением специальных одноязычных, бинарных и многоязычных словарей, предназначенных для машинного перевода.

Настоящая работа посвящена вопросам синтаксического анализа чешского текста и сопоставления синтаксических структур чешского и русского языков в целях машинного перевода.

В диссертации ставились следующие задачи:

1) выявление принципиальных возможностей и границ применения методов формального синтаксического анализа и необходимости введения элементов семантики в общую систему автоматического анализа чешского текста;

2) определение объема индивидуальной словарной информации (как грамматической, так и семантической), необходимого и достаточного для проведения автоматического синтаксического анализа чешского текста; разработка способов представления этой информации в одноязычном (чешском) и бинарном (чешско-русском) словарях, предназначенных для машинного перевода;

3) установление соответствий между элементарными синтаксическими конструкциями чешского и русского языков и построение формальных правил преобразования результатов независимого

синтаксического анализа одного языка в исходные данные для синтеза другого языка;

4) принципиальное решение проблемы снятия омонимии синтаксических конструкций (главным образом, предложных) в ходе автоматического синтаксического анализа чешского текста.

Диссертация состоит из трех глав.

I-я глава посвящена вопросам автоматического синтаксического анализа чешского текста.

Под синтаксическим анализом текста понимается установление отношений зависимости между словами текста.

Считается, что все слова текста на естественном языке связаны друг с другом по смыслу, в основном — попарно. Конкретные сочетания двух конкретных слов, объединенных по смыслу и связанных синтаксически, называются словосочетаниями. (Например: «дождь идет», «высокий человек», «после лекции», «быстро иду» и т. д.).

Если отвлечься от лексического значения компонентов словосочетаний, то нетрудно выделить несколько основных типов грамматической структуры словосочетаний (т. е. грамматических форм обоих членов сочетания, представленных в терминах классов слов, и типа синтаксической связи между ними).

Типовое сочетание представителей 2-х классов слов, имеющих определенную грамматическую структуру и связанных определенным типом синтаксической связи, мы называем конфигурацией. Один из членов конфигурации считается управляющим, второй — управляемым, или зависимым. Конфигурация принимается за основную синтаксическую единицу, используемую при анализе и сопоставлении синтаксических структур чешского и русского языков.

Для представления конфигураций в терминах классов слов предварительно были выделены основные классы чешских слов. Критерием для группировки слов в класс служило сходство функции и дистрибуции этих слов в тексте. Для чешского языка было выделено 13 таких классов слов (например: S — существительные; V — глаголы; A — прилагательные; M — вводные слова; Z — знаки препинания).

Далее были выделены грамматические различительные признаки (РП) чешского языка, в число которых включены как собственно грамматические признаки (например: 3 РП рода, 2 РП числа, 7 РП падежей, РП вида, наклонения; синтаксические РП сильного и слабого управления; РП отдельных предлогов и т. д.), так и семантические РП (например, признаки предметности, действия, места, времени, приблизительности, количества, результативности и т. п.). Всего было выделено 390 РП, из них 120 семантических РП.

В терминах классов слов и грамматических РП были записаны грамматические конфигурации чешского языка.

(Например: $S_7+V_{101\cdot16}$; $A_{52}+S$ и т. д.).

Всего было выделено 175 конфигураций, сгруппированных по типу синтаксического отношения, выражаемого членами каждой конфигурации, с учетом типа синтаксической связи между ними. Так были выделены классы конфигураций, выражающих основное предикативное отношение, вспомогательное внутрипредикативное отношение, служебное приглагольное, атрибутивное, объектное, релятивное и координационное отношения. В целях проведения машинного эксперимента по синтаксическому анализу чешского текста выделенные конфигурации были перегруппированы по типу управляющего класса, в результате чего в одну группу попадали все конфигурации, в которых в функции управляющего члена выступал один и тот же класс слов. Например:

V+S	или	S+A	или	F+S
V+D		S+N		F+F
V+M		S+S		F+E
V+L		S+F		$F+L_{63}$
V+V		S+D		
V+ ... и т. д.		S+ ... и т. д.		

Такой способ группировки конфигураций показывает потенциальную сочетаемость каждого класса слов, выступающего в функции управляющего члена соответствующих конфигураций, с другими классами слов. Потенциальные способности каждого класса слов (в функции управляющего) сочетаться с другими классами слов (в функции управляемых) называются активными валентностями*) каждого данного (управляющего) класса слов.

На основе перечней активных валентностей классов чешских слов была составлена грамматика валентностей чешского языка. Это было сделано потому, что для грамматики валентностей в Вычислительном центре ЛГУ Г. С. Цейтиным

*) Точное определение термина «валентность», а также общие принципы грамматики валентностей были разработаны Б. М. Лейкиной, С. Я. Фитиаловым, Г. С. Цейтиным. См., например: Лейкина Б. М., Некоторые аспекты характеристики валентностей, «Доклад на конференции по обработке информации, машинному переводу и автоматическому чтению текста», М., ВИНИТИ, 1961; Лейкина Б. М., Принципы построения английской грамматики для синтаксического анализа при автоматическом переводе, сборник «Научно-техническая информация», 1964, № 11, стр. 35—40; Фитиалов С. Я., О моделировании синтаксиса в структурной лингвистике, «Проблемы структурной лингвистики», М., 1964, стр. 100—114.

был составлен и запрограммирован Универсальный алгоритм синтаксического анализа, с помощью которого можно было провести эксперимент по автоматическому синтаксическому анализу чешского текста. Такой эксперимент был подготовлен и проведен в феврале 1963 г. совместно с Г. С. Цейтиным и С. Я. Фитиаловым в ВЦ ЛГУ на быстродействующей ЭЦВМ «Урал-1». В ходе эксперимента были использованы стандартные программы универсального алгоритма синтаксического анализа, составленные Г. С. Цейтиным.

Было проанализировано 63 чешских предложения, взятых из текстов научно-технического и политического характера. Каждое слово предложения было закодировано по восьмеричной системе, причем после информации о самом слове кодировалась информация о его левых и правых валентностях. Вся информация о тексте была пробита на перфоленту. В оперативную память машины вводились попеременно перфоленты с программами универсального алгоритма (УА) синтаксического анализа и перфоленты с закодированным текстом, разбитым на зоны (6 зон — не более 110 слов в каждой зоне). В результате работы 1-й программы (УА) на выходе печатающего устройства была получена лента с информацией о количестве вариантов анализа (для каждого предложения) и о зависимости каждого слова от других слов (в пределах предложения). После работы 2-й программы («конфигурационно-выделительной») на выходе печатались цифровые данные, представляющие все варианты анализа для каждого введенного предложения.

В общей сложности на анализирование 63 предложений (=650 слов) было затрачено около 10 часов машинного времени. В результате эксперимента в среднем на каждое предложение было получено от 2-х до 8-ми вариантов анализа, и лишь в отдельных случаях — 14, 35, 52, 192 и более вариантов.

После тщательной обработки результатов эксперимента были выявлены и записаны те необходимые «валентно-позиционные» условия, или ограничения, которые должны быть наложены на 1-й вариант грамматики валентностей чешского языка с целью получения правильных вариантов синтаксического анализа для каждого предложения чешского текста. Эти условия устанавливались следующим образом: анализировалась каждая неправильно замкнутая связь 2-х слов и определялись те необходимые условия, в результате проверки которых такая неправильная связь в принципе не могла бы быть установлена.

Далее был составлен 2-й вариант грамматики валентностей чешского языка, включающий как общие условия (например, условия обязательного согласования членов конфигураций в роде, числе, падеже, лице и т. д.), так и частные условия, или «условия на 2 слова» (т. е. активные валентности классов и подклассов

чешских слов), записанные в виде валентно-позиционных условий. В ходе составления и последующей домашинной проверки этой грамматики на чешских текстах научно-технического и политического характера было установлено, что для осуществления автоматического синтаксического анализа чешского текста необходимо учитывать не только общую синтаксическую информацию о классах и подклассах чешских слов (что обеспечивается грамматикой валентностей), но и индивидуальную синтаксическую и семантическую информацию от отдельных словах анализируемого текста, без обращения к которой автоматический анализ предложных и беспредложных именных конструкций, а также разрешение омонимии синтаксических конструкций представляются крайне затруднительными.

Для этой цели нами был составлен специальный словарь чешских слов и разработаны принципы введения индивидуальной синтаксической и семантической информации в словарные статьи этого словаря.

Индивидуальная синтаксическая информация включает в себя сведения о способности каждого знаменательного слова к управлению другими словами в тексте.

Мы различаем 2 основных типа управления: сильное и слабое.

Под сильным управлением понимается такая синтаксическая зависимость между словами текста, при которой управляющее слово предсказывает появление в тексте управляемого члена с достаточной степенью вероятности *) и всегда в определенной форме (например, в форме определенного падежа или определенного предлога).

В зависимости от силы управления, т. е. от степени вероятности совместной встречаемости в тексте управляющего слова и управляемой формы, нами выделяются 3 степени сильного управления (α, β, γ) и устанавливается иерархия этих степеней (α>>β>γ).

Например:

překládat	α	β	γ
(переводить)	*Akk.* (что)	*z* + *Gen.* (с чего)	*do* + *Gen* (на что)

Под слабым управлением нами понимается такое отношение зависимости между словами текста, при котором управляющее слово предсказывает лишь возможное, факультативное появление в тексте управляемого, зависимого члена, причем предсказывается определенное (как правило, обстоятельственное) значение этого зависимого члена, а не его форма.

*) Степень вероятности понимается здесь и далее не в количественном смысле, а лишь как степень предсказуемости управляемого члена со стороны управляющего.

Например, глагол «прийти» может предсказывать появление в тексте любой из следующих конструкций с обстоятельственным значением времени:

прийти { сегодня; во-время; зимой; в этот день; в четверг; перед грозой; после обеда; во время работы; под вечер; через час и др. }

Очевидно, что ни одна из этих конструкций не предсказывается глаголом «прийти» обязательно, с достаточной степенью вероятности. Здесь можно говорить уже о семантической валентности глагола «прийти» на целый класс конструкций, выражающих общее обстоятельственное значение времени.

Таким образом, в случае слабого управления между управляющим и управляемым членами существует, прежде всего, смысловая зависимость; управляющее слово предсказывает определенное значение управляемого члена, которое может быть выражено целым классом языковых конструкций, имеющих различную синтаксическую структуру.

Нами выделяется 5 степеней слабого управления (*a, b, c, d, e*), также различающихся по степени вероятности совместной встречаемости в тексте управляющего и управляемого членов ($a>b>c>d>e$).

Сведения о способности каждого знаменательного слова к сильному и слабому управлению записываются в словаре с помощью номеров соответствующих синтаксических РП, размещаемых в графах словарной информации, например:

dáti	α	β	γ	*a*	*b*	*c*	*d*	*e*
(дать)	201	243	—	210	211	215	214	224

где:

РП 201 — сильное управление 1-й степени винительным падежом;
РП 243 — сильное управление 2-й степени дательным падежом;

РП 210 —	способность к факультативному, слабому управлению классом конструкций	с обстоятельственным значением	места;
РП 211 —	»	»	времени;
РП 215 —	»	»	причины;
РП 214 —	»	»	цели;
РП 224 —	»	»	образа действия.

Очевидно, что в случае сильного управления — при условии правильного заполнения соответствующих граф словарной информации — автоматический синтаксический анализ не представляет принципиальных трудностей.

В случае же слабого управления установление смысловой зависимости между управляющим и управляемым членами предполагает предварительное знание значения управляемой формы, что весьма затруднительно в случае слабоуправляемых предложных конструкций, т. к.: 1) значения всех возможных предложных конструкций не могут быть заданы в словаре заранее; 2) значение каждой из этих конструкций не равно значению ни одного из ее компонентов, не есть сумма значений этих компонентов, а есть некоторое результирующее значение, которое может быть определено на содержательном уровне.

В целях предварительного автоматического определения значения каждой предложной конструкции чешского языка была составлена специальная таблица определения значений предложных конструкций на основе семантической информации о компонентах самих конструкций и о словах, управляющих этими конструкциями *). Такая таблица необходима, прежде всего, для определения значений слабоуправляемых предложных конструкций. Общий ход автоматического анализа каждой такой конструкции состоит в ее сопоставлении с таблицей, служащей своего рода трафаретом; в случае совпадения грамматической и семантической информации о компонентах конструкции и об управляющем слове со значениями, заданными в определенной строке таблицы, автоматически выводится значение анализируемой предложной конструкции, записанное в той же строке.

Таблица состоит из 36 разделов, соответствующих 36 наиболее употребительным чешским предлогам, каждый из которых управляет только одним падежом.

*) Предварительно каждому знаменательному слову и каждому предлогу в словаре была приписана индивидуальная семантическая и синтаксическая информация, представленная формальным образом — в виде комбинации соответствующих грамматических и семантических РП; значение каждой предложной конструкции, выводимое в ходе анализа, записывается аналогичным образом — в виде комбинации семантических РП.

В настоящее время таблица, словарь и грамматика валентностей чешского языка (2-й вариант) полностью подготовлены для проведения дальнейших экспериментов по синтаксическому анализу чешского текста, в результате которых предполагается разработать оптимальный вариант грамматики, обеспечивающей осуществление автоматического синтаксического анализа для любого вводимого в машину предложения чешского научно-технического текста.

II-я и III-я главы работы посвящены вопросам сопоставления синтаксических структур чешского и русского языков и построения формальных правил соответствия, что необходимо для этапа преобразования грамматической информации входного текста в грамматическую информацию выходного текста в ходе машинного перевода.

Эти вопросы, составляющие собственно перевод при машинном переводе, являются значительно менее разработанными, чем вопросы независимого анализа или синтеза отдельных языков.

Во **II-й главе** проводится сопоставление синтаксических структур чешского и русского языков на уровне элементарных синтаксических конструкций, или конфигураций. Количество и последовательность рассмотрения и сопоставления чешских и русских конфигураций были заданы составленной ранее таблицей грамматических конфигураций чешского языка.

В работе подробно рассматриваются лишь те конфигурации, грамматическая структура которых не совпадает в обоих языках. Результаты сопоставления несовпадающих конфигураций записываются в виде формальных правил соответствия, предназначенных для последующей алгоритмизации и использования в ходе автоматического перевода.

В классе конфигураций, выражающих основное предикативное отношение, все несовпадающие конфигурации могут быть разбиты на 2 группы:

1) конфигурации, которые различаются выражением субъекта; в этой группе рассматриваются конфигурации, в которых в функции субъекта выступает личное местоимение в форме именительного падежа, как правило, отсутствующее в чешских конфигурациях и регулярно присутствующее в русских, что обусловлено различием способов выражения грамматического значения лица в этих языках. (ср. *přijdeš* // ты придешь и т. д.);

2) конфигурации, которые различаются выражением предиката; здесь рассматриваются конфигурации, в которых в функции управляющего члена выступает глагол в отрицательной форме, выражаемой синтетическим способом в чешском языке и аналитическим — в русском (ср. *nečetl* // не читал и т. д.); глагол — связка в форме настоящего времени, присутствующий в чешском

языке и регулярно отсутствующий в русском (ср. *otec je doma* // отец дома и т. д.), а также частные конфигурации, в которых в функции управляющего члена выступают глагол *míti* // иметь

(ср. *mám knihu* // | у меня есть книга
я имею книгу

и безличный глагол (ср. *prší* // дождь идет и т. д.). В последнем случае несовпадения конкретных конструкций могут быть описаны лишь на словарном уровне — с помощью перекрестных ссылок в бинарном словаре.

В том случае, когда чешской конфигурации соответствует 2 и более русских варианта, такая вариантность фиксируется в русской части правила соответствия; выбор одного из нескольких возможных вариантов должен определяться специальными правилами русского синтеза.

В классах конфигураций, выражающих вспомогательное внутрипредикативное отношение (глагольное и глагольно-именное), рассматриваются, прежде всего, конфигурации, в которых в функции управляющего члена выступает глагольная форма прошедшего времени, а в функции управляемого — чешские формы настоящего времени вспомогательного глагола *býti*

(ср. *přišel jsem* // я пришел
и т. д.

Кроме того, здесь рассматриваются также конфигурации, управляющим членом которых является глагол-связка *býti* в форме настоящего времени, а управляемым — именная часть в форме именит. падежа

(ср. *tady je hezké* // здесь красиво
Václav je student // Вацлав — студент и т. д.)

а также ряд других несовпадающих конфигураций более частного характера.

В классе конфигураций, выражающих служебное (приглагольное) отношение, рассматриваются несовпадающие конфигурации, в которых аналитической форме возвратного глагола в чешском языке соответствует синтетическая форма возвратного русского глагола (ср. *vrátil se* // вернулся и т. д.), а также конфигурации, в которых в функции управляющего члена выступает глагольная форма прошедшего времени, а в функции управляемого в чешском языке — формы условного наклонения, служащие одновременно и для выражения грамматических значений лица и числа (*buch*, *bys*, *by*, *bychom*, *byste*), а в русском языке — частица «бы», являющаяся лишь показателем условного наклонения.

(ср. *přišli bychom* // мы бы пришли;
přišli byste // вы бы пришли и т. д.).

Далее, в классе конфигураций, выражающих атрибутивное отношение, наблюдается совпадение большей части конфигураций. Расхождения имеются лишь: 1) В тех случаях, когда в функции управляемого члена конфигураций выступают формы сравнительной и превосходной степеней качественных прилагательных, выраженные синтетически в чешском языке и аналитически — в русском

(ср.: *hlubší* // более глубокий
nejvroucnější // { наиболее горячий
самый горячий).

Единичные совпадения этих форм в чешском и русском языках фиксируются в бинарном словаре (ср. *lepší* // лучший; *horší* // худший и т. д.)

2) В группе конфигураций, выражающих кроме основного атрибутивного отношения дополнительное притяжательное отношение

(ср. *bratrův dům* // дом брата
Puškinova báseň // стихотворение Пушкина и т. д.)

В классе конфигураций, выражающих объектное отношение, чешско-русские соответствия принципиально не могут быть установлены на уровне конфигураций, т. к. объектное отношение реализуется с помощью сильного управления, которое всегда индивидуально, а потому может быть описано лишь на лексическом уровне — при условии составления специального бинарного словаря, в котором индивидуальная синтаксическая информация к каждой паре слов входного и выходного языков располагается по единому принципу с соблюдением строгого соответствия в заполнении граф сильного управления.

В работе предлагаются принципы заполнения граф словарной информации к каждой паре эквивалентных лексем входного и выходного языков в бинарном словаре, что иллюстрируется соответствующей таблицей.

Наибольшие трудности представляет сопоставление чешских и русских конфигураций, выражающих обстоятельственное отношение, и формальное описание соответствия управляемых членов этих конфигураций (т. е. слабоуправляемых конструкций). Прямой, непосредственный автоматический перевод таких конструкций с языка на язык является крайне затруднительным, а в ряде случаев просто невозможным. Значительно более надежным представляется «опосредствованный» перевод, или перевод через язык-посредник, в качестве которого может выступать набор элементарных ситуаций реальной действительности,

описываемых соответствующими слабоуправляемыми конструкциями («ситуационный язык»).

В работе предлагается принцип раздельного сопоставления таких конструкций, разбитых на классы в зависимости от выражаемых ими обстоятельственных значений.

III-я глава, состоящая из 2-х частей (А, Б), посвящена вопросам сопоставительного анализа русских и чешских слабоуправляемых конструкций, выражающих обстоятельственные значения места (часть А) и времени (часть Б).

Общий ход исследования в обеих частях следующий:

1) вначале составляется синтаксис реальных ситуаций, описываемых анализируемыми конструкциями;

2) затем проводится подробное сопоставление русских и чешских именных и отчасти — наречных конструкций, служащих для описания каждой из выделенных ситуаций;

3) поскольку в ряде случаев синтаксическая структура анализируемых конструкций определяется семантикой их компонентов, производится предварительная семантическая классификация *) слов, используемых в этих конструкциях, и слов, способных управлять этими конструкциями; при составлении семантической классификации учитываются не только значения слов, но и их «синтаксическое поведение» (в рамках анализируемых конструкций);

4) в результате сопоставления записываются формальные правила соответствия русских и чешских конструкций (главным образом, именных), учитывающие семантику компонентов этих конструкций, а в ряде случаев — и семантику управляющих слов.

В части А предварительно было выделено 50 элементарных пространственных ситуаций, сведенных в таблицу, где каждая из этих ситуаций была представлена схематически и снабжена содержательной и символической записью (в терминах семантических РП). Значение каждой элементарной пространственной ситуации мы называем простым местным значением (например, «внутри»), а значение нескольких элементарных ситуаций, объединенных общим характером взаимоотношения предметов в пространстве, мы называем сложным местным значением (например: «внутри — внутрь — изнутри»). В ходе детального сопоставления языковых конструкций, служащих для описания пространственных ситуаций (т. е. для выражения простых и сложных местных значений) в русском и чешском языках, была произведена семантическая классификация существительных, используемых в анализируемых конструкциях, и глаголов, способных управлять эти-

---

* Все семантические характеристики слов представляются формально — с помощью номеров соответствующих семантических РП.

ми конструкциями. В результате сопоставления было записано 80 формальных правил соответствия, предназначенных для последующей алгоритмизации и использования в ходе экспериментов по машинному переводу с русского языка на чешский и обратно.

В части Б проводится сопоставление русских и чешских конструкций, выражающих обстоятельственные временные значения. В отличие от части А здесь выделяются вначале сложные временные значения (всего — 9 таких значений), а затем в ходе сопоставления соответствующих языковых конструкций выявляются и уточняются простые временные значения (всего 37), каждое из которых выражается с помощью целого класса слабоуправляемых конструкций. Предлагается семантическая классификация существительных, используемых в этих конструкциях. В результате сопоставления было записано 70 формальных правил соответствия русских и чешских конструкций.

Следует отметить, что в обеих частях сопоставление конструкций с обстоятельственными значениями содержало элементы синтеза данных конструкций в чешском языке, т. к. каждой входной русской конструкции с обстоятельственным значением места или времени ставились в соответствие все возможные выходные чешские эквиваленты, представляющие собой синонимичные варианты с определенными стилистическими оттенками.

Поскольку пока мы не умеем еще даже содержательно определить условия, необходимые для выбора того или иного варианта, а тем более не можем формализовать эти условия, эти варианты задавались простым перечислением, причем на 1-е место в каждом таком перечне ставилась наиболее употребительная и нейтральная в стилистическом отношении конструкция; конструкции, для которых характерны существенные стилистические или какие-либо другие ограничения (напр., архаичные конструкции), заключались в скобки.

Кроме того, в ходе сопоставления нами решались вопросы многозначности (или — в более общем смысле — омонимии) предложных конструкций на основе семантического анализа контекста. Для разрешения многозначности каждой такой конструкции (ср., например: (ждать) «около часа» и (явиться) «около часа») в правилах соответствия задавались семантические РП слов, способных снять эту многозначность; чаще всего это были слова, управляющие данными конструкциями.

Введенный здесь принцип разрешения многозначности предложных конструкций на основе их «семантической дистрибуции» существенно отличается от метода снятия многозначности и омонимии на основе «лексической дистрибуции» омонимичных конструкций, применяемого в большинстве работ по машинному переводу. Различие заключается в том, что в 1-м случае задаются лишь определенные семантические РП, которые могут принадлежать большому числу слов, а во 2-м случае задаются перечни

конкретных слов, или лексем, входящих в состав контекстного окружения омонимичной конструкции и способных в большинстве случаев снять эту омонимию. Думается, что 1-й принцип, основанный на законах семантической сочетаемости слов в тексте, является более простым и надежным.

В **заключении** содержатся основные выводы и результаты проделанной работы, которые могут быть сведены к следующим пунктам:

1. Были разработаны общие принципы проведения автоматического синтаксического анализа чешского текста (1 глава).

В систему синтаксического анализа были введены элементы семантического анализа, что оказалось необходимым, прежде всего, для автоматического установления однозначной зависимости слабоуправляемых предложных конструкций от управляющих ими слов, а также для автоматического разрешения омонимии синтаксических конструкций.

Для этого предварительно были:

— выделены основные классы чешских слов (всего 13 классов);

— разработана сложная система грамматических и семантических различительных признаков (РП), релевантных для чешского языка (всего — 390 РП);

— в терминах классов слов и грамматических РП записаны грамматические конфигурации чешского языка (175 конфигураций);

—составлен 1-й вариант грамматики валентностей чешского языка в терминах классов чешских словоформ;

— подготовлен и проведен 1-ый эксперимент по автоматическому синтаксическому анализу чешского текста (в ВЦ ЛГУ на ЭЦВМ «УРАЛ-1»);

— после обработки результатов 1-го эксперимента составлен 2-ой вариант грамматики валентностей чешского языка с валентно-позиционными условиями, содержащей всю синтаксическую информацию о чешском тексте на уровне классов и подклассов чешских слов;

— установлено, что для автоматического синтаксического анализа предложных и беспредложных конструкций необходимо учитывать индивидуальную информацию о словах текста; для этой цели были составлены:

— специальный словарь чешских слов и разработаны принципы представления индивидуальной синтаксической и семантической информации к словам этого словаря;

— таблица определения значений слабоуправляемых предложных конструкций на основе семантической информации

о компонентах конструкций и о словах, способных управлять этими конструкциями.

2. В целях автоматического преобразования грамматической (главным образом, синтаксической) информации входного текста (здесь — чешского) в соответствующую информацию выходного текста (русского) в ходе машинного перевода было проведено предварительное сопоставление синтаксических структур чешского и русского языков на уровне элементарных синтаксических единиц, или конфигураций (II глава).

В результате такого сопоставления было установлено, что полное совпадение конфигураций чешского и русского языков наблюдается лишь в двух классах конфигураций, выражающих релятивное и координационное отношения.

В классах конфигураций, выражающих основное предикативное, вспомогательное внутрипредикативное, служебное приглагольное и атрибутивное отношения, наблюдается совпадение большей части конфигураций чешского и русского языков; все несовпадающие конфигурации этих классов могут быть описаны либо с помощью формальных правил соответствия (всего было записано 25 таких правил), предназначенных для последующей алгоритмизации, либо на словарном уровне — с помощью перекрестных ссылок в бинарном словаре, если несовпадение касается отдельных словосочетаний, входящих в соответствующие классы конфигураций.

В классе конфигураций, выражающих объектное отношение, сопоставление проводится на уровне бинарного словаря при условии соблюдения строгого соответствия в заполнении граф сильного управления для каждой пары знаменательных слов входного и выходного языков.

В классе конфигураций, выражающих обстоятельственное отношение, сопоставление может быть осуществлено лишь на семантическом уровне и в результате предварительного сопоставительного анализа управляемых членов этих конфигураций (т. е. слабоуправляемых конструкций).

3. В целях построения формальных правил соответствия слабоуправляемых конструкций, выражающих различные обстоятельственные значения в русском и чешском языках, был проведен подробный сопоставительный анализ таких конструкций, выражающих обстоятельственные значения места и времени, при условии соотнесения каждой из этих конструкций с соответствующей реальной ситуацией, описываемой данной конструкцией (III глава). Для этой цели предварительно были выделены элементарные и сложные пространственные и временные реальные ситуации (и соответствующие им простые и сложные обстоятельственные значения), совокупность которых может быть использована в качестве семантического языка-посредника, служащего для перевода

анализируемых конструкций с языка на язык. Для класса слабоуправляемых конструкций, выражающих обстоятельственные местные значения, было выделено 50 элементарных пространственных ситуаций (и соответственно столько же простых местных значений) и записано 80 формальных правил соответствия русских и чешских конструкций; для класса конструкций, выражающих обстоятельственные временные отношения, было выделено 37 простых временных значений и записано 70 правил соответствия русских и чешских конструкций.

В ходе сопоставительного анализа было выработано принципиальное решение проблемы снятия омонимии предложных конструкций на основе их «семантической дистрибуции», т. е. с учетом семантической информации слов, входящих в контекстное окружение этих конструкций.

Настоящая работа выполнялась во Всесоюзном институте научной и технической информации Государственного комитета по координации научно-исследовательских работ Совета Министров СССР и Академии наук СССР, частично — в Вычислительном центре Ленинградского университета, а также в Отделе алгебраической лингвистики и машинного перевода на философском факультете Карлова университета и в Отделе лексикологии и лексикографии Института чешского языка ЧСАН (г. Прага).

Основные теоретические положения работы и практические вопросы автоматического синтаксического анализа чешского текста были доложены и обсуждены на II-ой Всесоюзной конференции по автоматической обработке научной информации (Москва, ВИНИТИ, 14—16 мая 1963 г.) и на заседаниях Лингвистического объединения в Праге, Брно и Братиславе (июнь — июль 1964 г.)

Чешский материал диссертации и смысловые соответствия русских и чешских конструкций проверялись совместно с сотрудниками отдела алгебраической лингвистики и машинного перевода на философском факультете Карлова университета (г. Прага), частично — на кафедре русистики Брненского университета.

К работе дается приложение, включающее 12 таблиц.

СПИСОК

работ, опубликованных автором по теме диссертации

1. Пащенко Н. А. Некоторые вопросы автоматического синтаксического анализа чешского научно-технического текста, сборник «Научно-техническая информация», 1963, № 9.
2. Пащенко Н. А. Анализ и сопоставление способов выражения обстоятельственных временных значений в русском и чешском языках (в целях машинного перевода), ч. I, «The Prague Bulletin of Mathematical Linguistics», 1965, № 3, Praha
3. Пащенко Н. А. Анализ и сопоставление способов выражения обстоятельственных временных значений в русском и чешском языках (в целях машинного перевода), ч. II. «The Prague Bulletin of Mathematical Linguistics», 1965, № 4, Praha
4. Пащенко Н. А. К вопросу об автоматическом синтаксическом анализе предложных конструкций (на материале чешских и русских текстов) сборник «Научно-техническая информация», 1965, № 5.
5. Пащенко Н. А. О возможном формальном подходе к вопросу автоматического синтаксического анализа предложных и беспредложных именных конструкций чешского языка, «Kybernetika», Praha, 1965 (в печати).

В печать от 12/VIII — 1965 г. Заказ 4128 Тираж 200

Производственно-издательский комбинат ВИНИТИ,
Люберцы, Октябрьский пр., 403