N. A. Paščenko

# Об одном возможном подходе к вопросу автоматического синтаксического анализа предложных и беспредложных именных конструкций чешского языка

*Kybernetika*, Vol. 2 (1966), No. 1, (75)--85

Persistent URL: http://dml.cz/dmlcz/125330

KYBERNETIKA ČÍSLO 1, ROČNÍK 2/1966

# Об одном возможном подходе к вопросу автоматического синтаксического анализа предложных и беспредложных именных конструкций чешского языка*

Н. А. Пащенко

Настоящее сообщение содержит некоторые выводы, сделанные автором на основе работ по автоматическому синтаксическому анализу чешского текста и касающиеся вопросов автоматического установления синтаксической зависимости предложных и беспредложных именных конструкций от управляющих ими слов.

Предметом нашего рассмотрения являются вопросы автоматического синтаксического анализа предложных и беспредложных именных конструкций,** и прежде всего, вопросы автоматического установления однозначной зависимости этих конструкций от управляющих ими слов („слов-хозяев"), что представляет существенную трудность в ходе автоматического перевода с чешского языка.

Каждая из анализируемых конструкций выполняет в предложении определенную функцию и в соответствии с этим служит для выражения определенного значения (объектного или обстоятельственного); зависимость каждой такой конструкции от „слова-хозяина" синтаксически всегда реализуется путем управления.

Под *управлением* мы понимаем такой тип синтаксической зависимости между двумя словами текста, при которой одно слово (управляющее) предсказывает появление в тексте другого слова (управляемого) с некоторой степенью вероятности.***

* Настоящее сообщение содержит некоторые выводы, сделанные на основе работы по автоматическому синтаксическому анализу чешского текста. Фактическая часть работы заключена в диссертации автора, а также — частично — в статьях [1], [2].

** В понятие беспредложной именной конструкции нами не включаются именные формы именительного падежа, способные выполнять в предложении функцию субъекта.

*** Степень вероятности здесь и далее понимается нами как степень предсказуемости управляемого члена со стороны управляющего и определяется лишь интуитивно.

Нами различаются два основных типа управления: сильное и слабое.

Под *сильным управлением* понимается такой тип синтаксической зависимости, при котором управляющий член предсказывает появление в тексте управляемого, зависимого члена с достаточной степенью вероятности и всегда в определенной форме (например, в форме определенного косвенного падежа — в случае беспредложного управления, или в форме определенного, конкретного предлога — в случае предложного управления).

Поскольку сила управления, т. е. степень вероятности появления управляемого члена в зависимости от управляющего, может быть различной, целесообразно выделять степени силы управления. Опыт показывает, что ни одно знаменательное слово не может иметь при себе одновременно более трех сильноуправляемых членов. В соответствии с этим мы выделяем три степени сильного управления:

1-я степень — $\alpha$ — обязательное управление;
2-я степень — $\beta$ — менее вероятное появление управляемого в зависимости от управляющего — по сравнению с 1-й степенью;
3-я степень — $\gamma$ — еще менее вероятное появление управляемого в зависимости от управляющего — по сравнению с 1-й и 2-й степенями.

Эти степени сильного управления могут быть также названы сильными валентностями,* т.е. потенциальными способностями данного слова управлять той или иной формой с определенной силой.

Каждая степень сильного управления ($\alpha$, $\beta$, $\gamma$) является переменной величиной, способной принимать только 2 значения:

1. $\varkappa$ — сильное управление данным предлогом; здесь $\varkappa$ — переменная величина, которая может принимать значение сильного управления любым конкретным предлогом, управляющим, в свою очередь, только одним падежом (т. е. $\varkappa \rightarrow na_{\text{Akk.}} \vee na_{\text{Lok.}} \vee o_{\text{Akk.}} \vee o_{\text{Lok.}}$ и т. д.**).
2. $\delta$ — сильное управление беспредложной именной формой в данном косвеном падеже; здесь $\delta$ — переменная величина, способная принимать значение сильного управления любым из пяти косвенных падежей (т. е. $\delta \rightarrow$ Gen. $\vee$ Dat. $\vee$ Akk. $\vee$ Lok. $\vee$ Instr.).

Сведения о степени сильного управления и о значении этой степени записываются для каждого слова в соответствующих графах словарной информации таким образом, что в случае сильного управления 1-й степени ($C_\alpha$) в словарной информации к управляющему слову заполняется графа $\alpha$ ($\alpha = \varkappa \vee \delta$); в случае

* Термин „валентность“ в указанном выше смысле заимствован из работы Б. М. Лейкиной [3].

** Знак „$\vee$“ — знак дизъюнкции.

$C_\beta$, т. е. сильного управления 2-ой степени, управляющее слово имеет заполненной графу $\beta$ ($\beta = \varkappa \vee \delta$), причем $C_\beta$ уступает по силе управления $C_\alpha$, т. е.: $\beta < \alpha$; в свою очередь при $C_\gamma$ управляющее слово имеет заполненной графу $\gamma$ ($\gamma = \varkappa \vee \delta$); $C_\gamma$ уступает по силе управления и $C_\alpha$ и $C_\beta$, т. е. $\gamma < \beta < \alpha$.

Ср. например:

| | $\alpha = \varkappa$ | $\beta = 0$ | $\gamma = 0$ | |
|---|---|---|---|---|
| vměšovat se | $do_G$ | | | (vměšovat se *do* besedy) |

| | $\alpha = \delta$ | $\beta = \varkappa$ | $\gamma = 0$ | |
|---|---|---|---|---|
| zabalit | Akk | $do_G$ | | (zabalit knihu *do* papíru) |

| | $\alpha = \delta$ | $\beta = \varkappa$ | $\gamma = \varkappa$ | |
|---|---|---|---|---|
| překládat | Akk | $z_G$ | $do_G$ | (překládat text *z* ruštiny *do* češtiny) |

Под *слабым управлением* нами понимается такой тип зависимости, при которой управляющий член предсказывает лишь факультативное, возможное появление в тексте управляемого члена, причем предсказывается определенное *значение* этого управляемого члена, а не его форма. Это значение (как правило, обстоятельственное) может быть выражено целым классом конструкций, имеющих различную синтаксическую структуру; так например, в один класс могут входить наречные, предложные и беспредложные именные конструкции, а также придаточные предложения, выражающие общее обстоятельственное значение времени.

Поскольку в случае слабого управления предсказывается значение, а не форма управляемого члена, здесь можно уже говорить о смысловой, или семантической зависимости управляемого члена от управляющего. Мы выделяем 5 степеней слабого управления ($a$, $b$, $c$, $d$, $e$), которые также различаются по степени вероятности совместной встречаемости управляющего и управляемого членов в одном контексте (т. е. $a > b > c > d > e$) и могут быть названы слабыми, или семантическими валентностями управляющего слова.

Так например, в индивидуальной синтаксической информации к глаголу *překládat* кроме трех степеней сильного управления (см. выше, стр. 77) заполняются также 5 степеней слабого управления, а именно:

слабая валентность на класс конструкций, выражающих обстоятельственное значение:

$a$ — времени,
$b$ — места,
$c$ — цели,
$d$ — причины,
$e$ — образа действия.

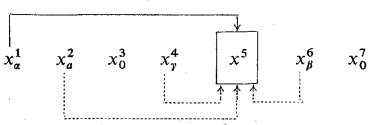
С учетом степеней слабого управления может быть записана следующая цепочка:

$$\alpha > \beta > \gamma > a > b > c > d > e\,.$$

Введение такой иерархии степеней управления необходимо для установления единственно правильной синтаксической связи в случае способности нескольких слов — в пределах одного и того же предложения — управлять одной и той же зависимой формой.*

Здесь вводится такое правило выбора управляющего слова для данного управляемого: если несколько слов одного предложения способны управлять данной зависимой формой (т. е. данным предлогом или данной беспредложной формой в функции управляемого) в счет различных степеней управления, то отношение зависимости этого управляемого устанавливается с тем управляющим словом, степень управления которого данной зависимой формой сильнее, чем у других потенциально возможных управляющих.

Например, в предложении из 7 слов:




1-е слово $(x_\alpha^1)$, 2-е слово $(x_a^2)$, 4-е $(x_\gamma^4)$ и 6-е слово $(x_\beta^6)$, способны управлять 5-м словом $(x^5)$, однако в счет разных по силе степеней управления.

В этом случае все гипотетические связи, изображенные пунктирными стрелками, снимаются в пользу самой „сильной" и единственно правильной связи,

* Сходные идеи мы находим в работах Л. Н. Иорданской [4], сотрудников группы математической лингвистики ВЦ ЛГУ [3], а также в работах лингвистической группы ВЦ Карлова университета в Праге [5].

изображенной сплошной стрелкой, исходящей от 1-го слова с различительным признаком (РП) сильного управления 1-й степени, т. е.:

$$x_{\alpha}^{1} \ \ldots \ x^{5} .$$

Формальные правила, обеспечивающие выбор наиболее „сильного" управляющего слова относительно каждого управляемого, записаны в виде валентно-позиционных условий во 2-м варианте грамматики валентностей чешского языка [1].

Сведения о наличии сильных и слабых валентностей у каждого данного слова, составляющие индивидуальную синтаксическую информацию, содержатся в словаре, где вся индивидуальная информация о слове закодирована в виде цифрового индекса при символе класса слов. Одно-, двух- и трехразрядные числа в индексе соответствуют определенным грамматическим и семантическим различительным признакам, введенным нами в анализирующую грамматику чешского языка. См. [1].

Например:

konfrontace — *S* 50.2.4.7.304.66.202.561 ,

где номера РП расшифровываются следующим образом:

50 — отглагольность
2 — женский род
4 — единственное число
7 — именительный падеж
304 — действие
66 — сравнение; сопоставление
202 — ($\delta$) — сильная валентность, соответствующая 1-ой степени сильного управления ($\alpha$) именной формой в родительном падеже (genitiv)
561 — ($\varkappa$) — сильная валентность 2-ой степени ($\beta$) на предлог „*S*", управляющий творительным падежом (instrumentál).

Аналогичным образом — в виде комбинаций соответствующих РП — записывается индивидуальная информация для всех остальных слов словаря.

В ходе автоматического синтаксического анализа предложных и беспредложных именных конструкций индивидуальная информация о словах текста используется различным образом, что определяется типом анализируемых конструкций, а именно: в случае анализа сильноуправляемых именных конструкций используется лишь индивидуальная синтаксическая информация, а в случае анализа слабоуправляемых конструкций привлекается еще и индивидуальная семантическая информация.

Общий ход автоматического анализа именных конструкций в пределах одного предложения следующий: вначале каждая анализируемая словоформа (предлог или именная форма в косвенном падеже), начиная с первой слово-

формы слева считается сильноуправляемой; для проверки этого предположения все остальные словоформы данного предложения проверяются на наличие РП сильного управления данной формой, т. е. на наличие РП = $\varkappa$ при анализе предлога и РП = $\delta$ при анализе беспредложной именной формы. В случае положительного ответа анализируемая словоформа считается сильноуправляемой; ей приписывается объектное значение (РП № 76) и устанавливается ее зависимость от управляющей словоформы, имеющей на нее сильную валентность.*

В случае же отрицательного ответа, т. е. в том случае, когда ни одна словоформа предложения не имеет сильной валентности на анализируемую словоформу, последняя считается слабоуправляемой, и для ее автоматиесчкой обработки уже необходимо привлечение индивидуальной семантической информации. Это обусловлено тем, что в случае слабого управления имеет место смысловая зависимость управляемой конструкции от управляющего слова. Очевидно, что для установления этой зависимости необходимо предварительное знание значения каждой слабоуправляемой конструкции. Выполнение этого условия является относительно простым в случае слабоуправляемой беспредложной именной конструкции, т. к. здесь семантическая информация к именной словоформе в большинстве случаев однозначно определяет ее вхождение в соответствующий класс слабоуправляемых конструкций. Например, если в результате анализа словоформа „dne" признана слабоуправляемой, то согласно основному семантическому РП ее индивидуальной информации (РП № 345 — „время") эта словоформа рассматривается как представитель класса слабоуправляемых конструкций, выражающих обстоятельственное значение времени ($К_{345}$). Затем проводится повторный анализ, в ходе которого все словоформы данного предложения проверяются на наличие семантической валентности на класс конструкций с обстоятельственным временным значением (РП 211), после чего — с учетом известных валентно-позиционных ограничений — замыкается связь, т. е. устанавливается однозначная зависимость анализируемой словоформы „dne" от управляющего слова.

Значительно более сложным является установление такой зависимости в случае предложных конструкций, т. к.:

1. значения всех возможных слабоуправляемых предложных конструкций принципально не могут быть заранее заданы в словаре;

2. значение каждой из этих конструкций не равно значению ни одного из ее компонентов, не есть сумма значений компонентов, а есть некоторое результирующее значение, которое может определяться, или выводиться из значений ее компонентов, а в ряде случаев — лишь с учетом значения одного или нескольких слов контекстного окружения предложной конструкции.

* Здесь, однако, действуют определенные валентно-позиционные ограничения, или условия, записанные в грамматике валентностей анализируемого языка [1].

В зависимости от характера значения этих конструкций и от сложности его определения мы выделяем три основных типа слабоуправляемых предложных конструкций:

А. Значение предложной конструкции определяется только значениями ее компонентов и не зависит от значений других элементов контекста.

Как правило, здесь в качестве предлога выступает вторичный предлог и притом не всякий, а лишь однозначный вторичный предлог.

Например: vstříc hostům, během týdne, díky tobě, následkem výbuchu, napříč ulicí, podél plotu, naproti domu, zpod koberce и т. д.

Б. Значение предложной конструкции определяется в зависимости от значений ее компонентов и от значения слова, управляющего этой конструкцией.

В предложных конструкциях этого типа, как правило, выступают первичные предлоги (v, na, s, po, u, při, z, za и др.) и — реже — неоднозначные вторичные предлоги (например: kolem, podle, přes и др.).

Определение значений таких конструкций сводится к разрешению омонимии предложных конструкций в условиях неомонимичного минимального контекста.

Например, неоднозначная чешская конструкция „na knihy" в разных контекстах реализует различные значения, каждое из которых в конечном счете определяется значением слова, управляющего этой конструкцией.

Ср.: 1. skříň *na knihy* (русский эквивалент — „шкаф для книг")
2. spadnout *na knihy* (русский эквивалент — „упасть на книги").

В соответствии с реализацией того или иного значения одна и та же конструкция получает различные семантические РП. Так в нашем примере предложная конструкция „na knihy" в 1-м случае получит РП № 350, что означает „назначение, предназначение", а во 2-м случае — РП 336.603, что означает „направление на поверхность предмета".

В. Значение предложной конструкции определяется не только на основе семантической информации о трех элементах (компонентах самой конструкции и слове, управляющем этой конструкцией), но и при условии привлечения дополнительной семантической информации еще об одном или нескольких элементах контекста.

Это случай омонимичности предложной конструкции в условиях омонимичного минимального контекста; такая омонимичность может быть разрешена лишь в условиях расширенного контекста.

В конструкциях этого типа могут быть использованы только первичные предлоги, характеризующиеся максимальной многозначностью. Однако рассмотрение таких предложных конструкций выходит за рамки нашей работы и может явиться предметом специального исследования.

* * *

В целях осуществления автоматического синтаксического анализа слабоуправляемых предложных конструкций первых двух типов (А, Б) наши была составлена таблица выведения значений этих конструкций из значений их компонентов (в случае типа А) а в ряде случаев с учетом значений слов, управляющих этими конструкциями (тип Б).

Каждая строка таблицы представляет собой правило определения значения предложной конструкции ($K^F$), которое может быть записано в виде импликации.

Предварительно обозначим индивидуальную информацию о предлоге ($F$), имени в форме косвенного падежа ($S$) и управляющем слове ($У$), записанную в виде комбинаций соответствующих РП, через:

$П$ – комбинация РП при данном предлоге ($F_П$);
$И$ – комбинация РП при данном имени существительном ($S_И$);
$Г$ – комбинация РП при данной словоформе (как правило, глагольной) управляющей данным предлогом ($У_Г$).*

Тогда в случае анализа слабоуправляемой предложной конструкции типа А правило определения ее значения (X) может быть записано следующим образом:

$(П\,.\,И) \rightarrow X$, что читается так: „Если информация ($П$) к предлогу ($F_П$) имеет данное значение, указанное в данной строке таблицы, и информация ($И$) к именной форме ($S_И$) также удовлетворяет заданному в этой строке значению, то значение всей предложной конструкции ($K^F$) есть переменная $X$, принимающая здесь данное обстоятельственное значение, записанное формально в виде комбинации семантических РП, содержащихся в соответствующих графах данной строки таблицы".

В случае слабоуправляемой предложной конструикции типа Б правило определения ее значения ($Y$) принимает вид:

$(Г\,.\,П\,.\,И) \rightarrow Y$, что читается так: „Если информация ($Г$) к управляющему слову имеет данное значение, соответствующее формальной записи в виде комбинации семантических РП в данной строке таблицы, и информация ($П$) к предлогу, как и информация ($И$) к именной форме, также удовлетворяют заданным значениям, записанным в данной строке, то значение всей предложной конструкции ($K^F$) есть переменная $Y$, принимающая здесь данное значение, содержащееся в соответствующих графах той же строки таблицы и также записанное в виде комбинации семантических РП".

В рассматриваемой таблице все заданные типы слабоуправляемых предложных конструкций сгруппированы не по выражаемому ими значению, а по пред-

* Разумеется, в функции управляющей словоформы может выступать не только глагольная, но и именная словоформа, образованная от имени существительного или прилагательного.

логу, т. е. в одну группу входят все конструкции, использующие один и тот же предлог и выражающие разные значения.* Такой принцип группировки облегчает поиск соответствующей строки в таблице.

В качестве иллюстрации мы рассмотрим здесь 3 строки нашей таблицы, отведенные для слабоуправляемых конструкций с предлогом „podle".

| № строки | РП управляющего слова | | РП предлога | | РП управляемого | | РП $K$ | Примеры |
|---|---|---|---|---|---|---|---|---|
| | грамм. РП | семант. РП | грамм. РП | семант. РП | грамм. РП | семант. РП | семант. РП | |
| 142 | 210 | 301/304/342 | 136 | 344 | 10 | 301 | 344 . 605 | stát podle stolu |
| 143 | 236 | 305 . 366 | | 376 | | 301 | 376 . 605 | jít podle trati |
| 144 | 220 | | | 332 | | $\overline{301}$ | 332 | podle úmluvy, podle mínění |

Очевидно, что строки № 142 и 143 представляют собой правила определения значений предложных конструкций типа Б (т. к. здесь учитывается семантическая информация об управляющем слове), а строка № 144 — правило определения значения слабоуправляемой конструкции типа А (в этом случае значение управляющего слова не существенно). РП, использованные в этих строках таблицы, расшифровываются следующим образом:

210 — слабая валентность на класс конструкций, выражающих обстоятельственное значение места ($K_{344}$)
236 — слабая валентность на класс конструкций, выражающих обстоятельственное значение направления — перемещения определенным путем — без указания цели ($K_{376}$)
220 — слабая валентность на класс конструкций, выражающих обстоятельственное значение относительного соответствия ($K_{322}$)
301 — предмет; $\overline{301}$ — не-предмет
304 — действие; 342 — состояние
305 — движение
366 — потенциальная направленность
136 — предлог „podle", управляющий Genitiv-ом
10 — Genitiv
344 — место в пространстве
376 — направление перемещения определенным путем (без указания цели)
605 — нахождение предмета возле чего-л., рядом с чем-л.
332 — соответствие (относительное)

* В настоящее время таблица содержит 196 строк и состоит из 36 разделов, соответствующих 36 наиболее употребительным чешским предлогам, каждый из которых управляет только одним падежом.

Таким образом, общий ход автоматического анализа слабоуправляемых предложных конструкций состоит в сопоставлении каждой такой конструкции, выявленной на предыдущем этапе анализа, с определенным трафаретом — таблицей выведения значений этих конструкций, в результате чего устанавливается точное значение каждой конструкции и управляемость этой конструкцией со стороны семантически и грамматически определенной управляющей формы.

Затем после повторного анализа предложения синтаксическая зависимость каждой данной предложной конструкции устанавливается с ближайшей словоформой (с известными валентно-позиционными органичениями), имеющей семантическую валентность на данное обстоятельственное значение, выражаемое анализируемой слабоуправляемой конструкцией.

В заключение следует сказать, что правила определения значений слабоуправляемых предложных конструкций основаны на законах семантической сочетаемости слов в тексте, которые существуют объекгивно и могут быть открыты исследователем в ходе синтаксического и семантического анализа текстов на данном языке с применением методов дистрибутивного и трансформационного анализов.

Дальнейшие исследования в этой области представляются нам весьма перспективными; очевидно, что результаты этих исследований могут быть использованы для построения семантического языка-посредника, предназначенного для автоматического перевода.

Автор пользуется случаем и выражает глубокую благодарность Вяч. Вс. Иванову и П. Новаку (Прага) за целый ряд ценных замечаний и предложений, высказанных ими по поводу данной работы.

(Поступило 5. марта 1965 г.)

ЛИТЕРАТУРА

[1] Пащенко Н. А.: Некоторые вопросы автоматического синтаксического анализа чешского научно-технического текста. „Научно-техническая информация“ (1963), № 9, стр. 38—43.

[2] Пащенко Н. А.: Анализ и сопоставление способов выражения обстоятельственных временных значений в русском и чешском языках (в целях машинного перевода), ч. I, II. „The Prague Bulletin of Mathematical Linguistics“, Praha 1965, № 3—4.

[3] Лейкина Б. М.: Некоторые аспекты характеристики валентностей. Докл. на конференции по обработке информации, машинному переводу и автоматическому чтению текста, ВИНИТИ, Москва 1961.

[4] Иорданская Л. Н.: Два оператора обработки словосочетаний с „сильным управлением“ (для автоматического синтаксического анализа). ИЯ АН СССР, Москва, 1961.

[5] Sgall P., Panevová J., Pitha P., Pala K.: Ze syntaktické analýzy češtiny (Přípravné práce ke SP). „AUC Slavia Pragensia“ III (1961), № 3, 181—196.

# Příspěvek k automatické syntaktické analýze předložkových a bezpředložkových nominálních konstrukcí v češtině

N. A. PAŠČENKO

Práce je věnována otázkám automatické syntaktické analýzy českých předložkových a nepřímých pádů. Závislost každého z těchto pádů na členu řídícím se vždycky realizuje rekcí. Rozeznáváme 2 základní typy rekce (silnou a slabou) a mimo to 8 stupňů rekce, které se rozlišují podle míry těsnosti vztahu mezi členem řídícím a závislým.

Popisuje se celkový postup automatického zjišťování syntaktické závislosti jakéhokoliv předložkového a nepřímého pádu na řídícím členu.

Nejobtížnější je analýza slabě řízených předložkových pádů, k čemuž byla sestavena speciální tabulka vyvozování významů předložkových vazeb na základě sémantické a gramatické informace o jejich komponentech a o slovech řídících. Během automatické syntaktické analýzy českého textu se každý předložkový pád srovnává s touto tabulkou a tím se zjišťuje jeho význam a jeho závislost na slově řídícím, které je v tabulce určeno po stránce sémantické a gramatické.

*Н. А. Пащенко, ВИНИТИ НМО, Балтийская ул. 14, Москва А 219, СССР.*