Petr Sgall

Matematický popis přirozených jazyků

Pokroky matematiky, fyziky a astronomie, Vol. 23 (1978), No. 3, 140--148

Persistent URL: http://dml.cz/dmlcz/139929

Matematický popis přirozených jazyků*)

(K 25. výročí MFF UK)

Petr Sgall, Praha

1. Teorie automatů patří k těm matematickým disciplínám, na jejichž vývoji je nejlépe vidět, jak se dnes prolínají matematické problémy s lingvistickými a jak je spolupráce těchto oborů důležitá nejen pro lingvistiku, v níž se metody kvantitativního a zejména algebraického zpracování rozvíjejí stále šíře a účinněji, ale také pro samu matematiku a pro společné matematicko-lingvistické aplikace v oblasti informatiky v nejširším slova smyslu (od programování po automatickou selekci informací a komunikaci člověka se strojem). Práce NOAMA CHOMSKÉHO [1, 2, 3], v nichž se gramatika — procedura vymezující množinu řetězů interpretovaných jako správně tvořené věty daného jazyka (přirozeného nebo umělého) — chápe jako formální systém (generující takovou množinu řetězů), vedly k vytvoření stupnice gramatik, která je už od svého vzniku v padesátých letech velmi důležitá i pro teorii automatů jako matematickou disciplínu a pro teorii programovacích jazyků.

Zastavme se nejprve u některých důležitých pojmů, jejichž definice je třeba uvést:

Řetězem nad abecedou (slovníkem) A nazýváme konečnou posloupnost $x_1, \ldots, x_k$, kde platí, že x_i je prvkem A pro $1 \leqq i \leqq k$; mluvíme i o řetězu prázdném, který neobsahuje žádný symbol. Řetězy se zpravidla zapisují v podobě $x_1 \ldots x_k$, tedy bez čárek mezi symboly.

Sřetězení čili konkatenace x a y je operace, která dvěma řetězům, x a y, přiřazuje řetěz xy; je-li $x = x_1 \ldots x_k$, $y = y_1 \ldots y_m$, je řetěz xy definován jako $x_1 \ldots x_k y_1 \ldots y_m$.

Volným monoidem rozumíme množinu všech řetězů (včetně řetězu prázdného) nad danou abecedou; volný monoid nad abecedou A označujeme A^*.

Formální jazyk L je množina řetězů nad abecedou A (tedy $L \subseteq A^*$).

Generativní gramatika je čtveřice $G = (V, V_T, P, S)$, kde V je konečná množina symbolů (abeceda nebo slovník), V_T je podmnožinou V (je to tzv. terminální slovník), P je konečná množina dvojic (nazývaných pravidly) tvaru (u, v), kde $u \in (V - V_T)^*$, $v \in V^*$, přičemž u je řetěz neprázdný (obvykle se však tato tzv. přepisovací pravidla zapisují v podobě $u \to v$, což čteme „přepiš u na v"), a $S \in V - V_T$ (S nazýváme výchozím symbolem, $V - V_T$ je množina symbolů pomocných).

Derivace je posloupnost řetězů, která vyhovuje těmto podmínkám: budiž dána generativní gramatika $G = (V, V_T, P, S)$ a řetězy $w, y \in V^*$; píšeme $w \Rightarrow y$, existují-li z_1, z_2, u a v tak, že $w = z_1 u z_2$, $y = z_1 v z_2$ a $(u, v) \in P$; posloupnost řetězů $w_0, \ldots, w_r$ nazýváme derivací, jestliže pro $0 \leqq i \leqq r - 1$ platí $w_i \Rightarrow w_{i+1}$ nebo jestliže $w_0 = w_r$.

*) Za cenné připomínky k textu této stati jsem vděčen prof. dr. M. NOVOTNÉMU, DrSc.

Jazykem generovaným gramatikou $G = (V, V_T, P, S)$ nazýváme množinu $L(G) = \{x \in V_T^*;$ existuje derivace tvaru $S, \ldots, x\}$.

Jako dva často rozbírané specifické typy generativních gramatik uvedeme tzv. (frázovou) gramatiku kontextovou a nekontextovou.

Generativní gramatika $G = (V, V_T, P, S)$ je kontextová, jestliže každé její pravidlo má tvar $uzv \to uyv$, kde $z \in V - V_T$; $u, v \in (V - V_T)^*$ a y je neprázdný prvek V^*. Jazyk generovaný kontextovou gramatikou je kontextový jazyk.

Nekontextová gramatika, na kterou se obvykle soustřeďuje nejvíce pozornosti, bývá definována jako taková generativní gramatika $G = (V, V_T, P, S)$, jejíž každé pravidlo má tvar $z \to v$, kde $z \in V - V_T$ a $v \in V^*$. Jazyk generovaný nekontextovou gramatikou je nekontextový jazyk.

Z oblasti přirozeného jazyka lze jako jednoduchý příklad uvést nekontextovou gramatiku s těmito pravidly (která ukazují i složení slovníku); *NP* tu čteme jako „nominální fráze", tj. podstatné jméno a jeho rozvití, *VP* je „verbální (slovesná) fráze", popř. „přísudková část" věty (každou z uspořádaných dvojic *NP*, *VP* považujeme za jediný symbol):

$S \to NP\ VP$
$NP \to A\ N$
$NP \to N$
$VP \to V\ NP$
$V \to$ *čte*
$N \to$ *otec*
$N \to$ *dopis*
$A \to$ *starý*

K terminálním řetězům tu patří *Otec čte dopis*, *Otec čte starý dopis* atd.; bylo by ovšem generováno také např. *Starý dopis čte dopis.*

Pro popis složité soustavy přirozeného jazyka nekontextová gramatika (ani po přidání mnoha pravidel zachycujících mluvnické jevy daleko jemněji, než jak je tomu v našem příkladu) nestačí. O významu tohoto typu gramatik však svědčí několik okolností:

(a) V celé stupnici gramatik se dosud nenašel typ, který by mohl generovat množinu řetězů odpovídající gramaticky správným větám přirozeného jazyka a který by přitom nebyl příliš silný. Gramatiky kontextové jsou obecnější než nekontextové, a tedy silnější, tzn. mohou generovat i jazyky, které nelze generovat žádnou nekontextovou gramatikou (např. množinu všech řetězů tvaru xx). Jejich nevýhodou však je, že neukazují specifičnost přirozeného jazyka jako takového a nelze na ně přenést dobře propracovanou teorii nekontextových jazyků. Regulární gramatiky (s pravidly tvaru $A \to aB$, kde a je symbol terminální, A a B jsou pomocné) jsou ovšem slabší než nekontextové; nelze jimi generovat ani tzv. jazyk zrcadlového odrazu, tj. množinu řetězů tvaru $x\hat{x}$, kde $\hat{x}$ se liší od x jen opačným uspořádáním symbolů. Ani ostatní typy dosud do stupnice gramatik zařazené nejsou pro popis přirozených jazyků vhodné. Hledají se proto lepší možnosti v kombinování několika systémů za sebou, přičemž (jak uvidíme v odd. 2) první složkou takového složeného aparátu bývá právě nekontextová gramatika nebo gramatika typu

s ní ekvivalentního; dva typy generativních gramatik se považují za ekvivalentní právě tehdy, jestliže každý jazyk generovaný některou gramatikou typu prvního je generován také některou gramatikou typu druhého a naopak.

(b) Řada lingvistických postupů, kterých se tradičně užívalo v nejrůznějších školách pro neformální popis přirozených jazyků, byla, jak souhrnně ukázal POSTAL [4], při explicitní formulaci charakterizována jako ekvivalentní s nekontextovou gramatikou. Jde o takové systémy, jako tzv. metoda bezprostředních složek (známá z lingvistiky anglosaských zemí, u nás uplatňovaná v tradiční lingvistice jen zčásti, např. při dělení věty na část podmětovou a přísudkovou), blízká závorkování běžnému ve formálních jazycích, i tzv. závislostní syntax (při níž se mluví o členu řídícím — např. slovesu — a o členech na něm závislých, jako je předmět nebo příslovečné určení, tzn. členy, které jsou co do svého výskytu a popř. i co do svého tvaru — např. pádového — determinovány slovesem), zachycující strukturu věty v podobě určitého typu stromu (s vrcholem), ale patří sem i řada systémů vytvořených v rámci příprav strojového překladu aj.*)

(c) Jak ukázal zejména CHOMSKY [2] — viz též BAR-HILLEL [5] a GROSS [6] — jednotlivé typy gramatik jsou ve výše uvedeném smyslu ekvivalentní i s jednotlivými typy abstraktních automatů, jak jsou matematicky zachyceny např. u MCNAUGHTONA [7]. Nemůžeme se tu těmito vztahy zabývat podrobněji, ale všimneme si alespoň jednoho bodu z této bohaté problematiky, totiž ekvivalence mezi nekontextovými gramatikami a zásobníkovými automaty (k jejichž uplatnění v popisu přirozeného jazyka se vrátíme v odd. 2).

Zásobníkový automat (z. převodník) lze definovat jako pětici $T = (V_1, V_2, V_0, S, P)$, kde V_1, V_2, V_0, a S jsou konečné neprázdné množiny symbolů, přičemž S obsahuje s_0, a P je množina dvojic tvaru $((c, d, s), (s', d', c'))$ takových, že řetěz $cdd'c'$ je neprázdný; s a s' jsou prvky S; d a d' jsou prvky V_2^*; c a c' jsou buď prázdné, nebo prvky po řadě V_1 a V_0. Množiny V_1, V_2 a V_0 jsou interpretovány jako vstupní, zásobníková a výstupní abeceda (tedy množiny přípustných symbolů pro jednotlivé pásky, z nichž zásobníková páska slouží u tohoto typu automatu jako paměťová), S je množina vnitřních stavů, s_0 je stav počáteční a koncový, a P je definující funkce. Definující funkce automatu je v podstatě program jeho činnosti neboli předpis, podle kterého se na základě informace ze vstupní a zásobníkové pásky střídají vnitřní stavy a provádějí se operace na všech páskách.

Dále popíšeme proces zpracování vstupního řetězu, který nazveme komputací.

Nechť $T = (V_1, V_2, V_0, S, P)$ je zásobníkový převodník; komputace tohoto převodníku je posloupnost dvojic $(p_1, e_1), \ldots, (p_n, e_n)$, $n \geqq 1$, splňující tyto podmínky:

(1) $p_i = ((c_i, d_i, s_i), (s_i', d_i', c_i')) \in P$ pro $i = 1, 2, \ldots, n$;

(2) $s_i = s_0$ právě když $i = 1$; $s_i' = s_0$ právě když $i = n$;

---

*) Sem se řadí i kategoriální gramatiky, zpracované původně BAR-HILLELEM (na základě podnětů AJDUKIEWICZOVÝCH) a známé dnes zejména ze sémanticky orientovaných prací MONTAGUOVÝCH a D. LEWISE, které ukazují, že v oblasti sémantiky dochází dnes k ještě těsnější součinnosti matematiky (a logiky) s lingvistikou než v syntaxi. Zdaleka tu nejde jen o jednosměrné aplikace. (U nás viz k tomuto okruhu otázek [18], [19].)

(3) $s_{j+1} = s'_j, j = 1, 2, \ldots, n - 1$;

(4) e_1 je prázdný řetěz;

(5) existuje $e'_j \in V_2^*$ takové, že $e_j = e'_j d_j$ pro $j = 1, 2, \ldots, n$; $e_{j+1} = e'_j d'_j$ pro $j = 1, 2, \ldots, n - 1$; e_n a d'_n jsou prázdné řetězy.

Řetěz $c_1 \ldots c_n$ ($c'_1 \ldots c'_n$) nazveme vstupním (výstupním) řetězem komputace k.

Vstupní (výstupní) jazyk ${}^iL(T)$ (${}^oL(T)$) zásobníkového převodníku T je množina všech vstupních (výstupních) řetězů všech jeho komputací.

Na začátku je zásobníkový převodník v počátečním stavu s_0, čtecí hlava je na prvním symbolu vstupního řetězu, zásobník a výstupní páska jsou prázdné. Na konci komputace je přečten celý vstupní řetěz, zásobník je prázdný a převodník je ve stavu s_0.

Jak ukázal Chomsky, třída všech jazyků nekontextových je identická s třídou všech vstupních jazyků zásobníkových automatů, a také s třídou všech jejich jazyků výstupních; v tomto smyslu lze tedy mluvit o ekvivalenci nekontextových gramatik a zásobníkových automatů.

Poznamenejme ještě, že pro teorii programovacích jazyků mají dnes velmi značný význam nejen Chomského gramatiky (literatura o nekontextových jazycích je už dnes velmi rozsáhlá: programovací jazyky mají sice — podobně jako jazyky přirozené — určité vlastnosti, které se vymykají z nekontextových gramatik, ale jsou to zpravidla jen vlastnosti okrajové), ale také gramatiky závislostní (srov. např. SGALL [8], GORALČÍKOVÁ [9]).

2. Jak jsme se už zmínili, má-li být popis přirozeného jazyka formulován jako korektní matematický systém (což je předpokladem pro sestavení co nejúspornějších algoritmů a počítačových programů pro automatické zpracování textu, dialog člověka se strojem ap., viz odd. 3), je třeba pracovat nejen s gramatikou, ale se složitějším systémem, který gramatiku spojuje s dalšími složkami popisu.*) Obvykle se předpokládá, že gramatika (nekontextová nebo s ní ekvivalentní) generuje zápisy vět, které se dost liší od obvyklé vnější formy vět přirozeného jazyka, a tyto zápisy postupně zpracovává několik dalších zařízení (transformační složka, popř. automaty typu převodníku, které řetězy svého vstupního jazyka překládají na řetězy jazyka výstupního). Aniž bychom zabíhali do složité empirické problematiky lingvistické, připomeňme jen, že je obvyklé rozlišování alespoň tří rovin, totiž fonologické (hláskové, při dosavadních praktických aplikacích ovšem nahrazované pravopisným zápisem), morfologické (s jednotkami, jako jsou pády, čísla, osoby, časy, způsoby a ovšem kmeny slov) nebo i povrchově syntaktické (s členěním věty na přísudkové sloveso, jeho podmět, předmět, příslovečné určení místa, času, způsobu atd.) a hloubkově syntaktické (s konatelem, adresátem, nástrojem atd., kde je zachycena významová shoda činných a trpných tvarů slovesa, dále vedlejší věty a její tzv. zkrácené podoby s infinitivem, přechodníkem, slovesným jménem ap.); po-

*) Existuje i jiný přístup k matematickému zvládnutí lingvistické problematiky, totiž tzv. analytické modely, které nekladou důraz na existenci efektivních procedur, nýbrž algebraickými metodami zpracovávají jednotlivé základní pojmy; k předním odborníkům tohoto směru patří rumunský matematik MARCUS, u nás prof. NOVOTNÝ, viz jeho přehlednou stať [20] a práce tam uvedené.

sledně uvedená rovina bývá někdy ztotožňována s rovinou významových zápisů vět, jindy se o významových zápisech mluví teprve při uplatnění formálního logického jazyka (obvykle jde o více nebo méně propracovanou obdobu predikátového kalkulu kromě některých nových prací, srov. pozn. na str. 142). Rovinou zde — s určitým zjednodušením — rozumíme množinu zápisů vět užívajících stejného aparátu (jednotky a relace mezi nimi, popř. operace vytvářející z jednotek elementárních jednotky komplexní, odpovídající slovním tvarům, skupinám slov a větám), přičemž každá věta má na každé rovině alespoň jeden zápis (o více zápisů půjde při víceznačnosti věty). Podrobnější přehled lingvistické problematiky i různých přístupů k formálnímu zpracování je podán v [10].

Nejrozšířenějším takovým systémem o několika složkách je transformační gramatika, popsaná původně u Chomského [1]; v jejím dalším vývoji několikrát došlo k zásadnímu přehodnocení vztahu mezi jednotlivými složkami a rovinami, a nyní je Chomského škola rozštěpena na dva protichůdné směry. Podle Chomského a jeho přívržence generuje gramatika (první složka, nekontextová) základní schéma větné hloubkové struktury, složka transformační doplňuje konkrétní lexikální obsazení těchto struktur a dále je převádí na struktury povrchové, které pak konečný převodník překládá na hláskové řetězy. Popis sémantiky je tu chápán jako více méně paralelní s transformační složkou, tj. interpretačními pravidly se přiřazuje význam jednotlivým hloubkovým strukturám; značné potíže jsou však způsobeny tím, že v mnoha případech jsou věty lišící se při tomto pojetí jen v povrchové struktuře přesto významově různé (srov. např. „Mnoho lidí uslyšelo výbuch" a „Výbuch uslyšelo mnoho lidí"); je tedy nutné i rozdílům intonačním a slovosledným přiřazovat sémantickou interpretaci, což situaci velmi komplikuje.

To je právě jedním z hlavních argumentů odštěpenců, jako je G. Lakoff a J. D. McCawley (stati charakterizující tuto diskusi a celkovou dnešní situaci v transformační gramatice vyšly česky v [11]). Ti pracují s výchozí složkou generující sémantické zápisy vět, převáděné potom transformační složkou atd. na jednotlivé nižší roviny. Po stránce matematické je ovšem tento přístup (tzv. generativní sémantika) zatím málo propracován a je mu právem vytýkáno, že počítá s tzv. globálními omezeními (k jejichž uplatnění je třeba mít k dispozici celou posloupnost transformací uplatněných v rámci popisu dané věty), která jsou příliš silným aparátem; nemá totiž velkou poznávací hodnotu, zachycuje-li se struktura přirozeného jazyka aparátem, který může vymezit jakoukoli rekurzívně spočetnou množinu řetězů a neukazuje tedy nic specifického o přirozeném jazyce, neříká nic o jeho základních vlastnostech a nedovoluje dospět k matematicky zajímavým výsledkům. Přes řadu pokusů nejsou dosud ani jiné varianty transformační gramatiky spojeny s aparátem, který by těmto požadavkům vyhovoval.

Tím cennější je, že se podařilo ukázat, že popis jazyka, který už od první poloviny šedesátých let postupně formuluje skupina matematických lingvistů na Univerzitě Karlově (nyní na její matematicko-fyzikální fakultě) požadované vlastnosti má, neboť — jak ve stručnosti dále uvedeme — zachycuje přirozený jazyk jako systém specifického typu [12].*) Jde o popis založený už od počátku na podobném pohledu na vztah rovin jazykové struktury (jako lineárně uspořádaných od významu k hláskové podobě), s jakým

více méně současně vystoupila i řada jiných skupin matematické lingvistiky, zejména v SSSR (sem patří i práce leningradské skupiny CEJTINOVY, i BELECKÉHO, GLADKÉHO a dalších sovětských matematiků), ale i v USA, ve Francii aj., a jaký byl později přijat v rámci transformační gramatiky s výše uvedeným přístupem generativní sémantiky. Pražský přístup pracuje s nekontextovou (nebo závislostní) gramatikou generující významové zápisy vět a se čtyřmi zásobníkovými převodníky operujícími postupně za sebou, tzn. převádějícími významové zápisy (viz ukázku v tab. 1) nejprve na rovinu povrchové syntaxe, pak na rovinu morfologickou (z níž jsou zápisy konečným automatem převedeny do hláskové nebo grafické podoby). Jak bylo ukázáno v uvedené stati M. PLÁTKA [12], každý další takto připojený zásobníkový převodník může rozšířit třídu generovaných jazyků. (Jde o zásobníkové automaty vyhovující určitým specifickým omezením co do možnosti zkracovat zpracovávané řetězy aj.)

Ilustrujme problém příkladem: Gramatika G_1 generuje množinu řetězů $hxc\hat{x}h$, tj. jazyk zrcadlového odrazu, nad abecedou se dvěma symboly a, b, jen s tím rozdílem, že konec řetězu x je tu označen symbolem c a hranice celého řetězu symbolem h. Takto generovaný jazyk je podmnožinou vstupního jazyka zásobníkového automatu P_1, který převádí jednotlivé řetězy na prvky množiny řetězů tvaru $hxxh$. Posledně uvedená množina řetězů (jak jsme uvedli, je to kontextový jazyk, pro který neexistuje nekontextová gramatika) je tedy generována systémem o dvou složkách: nekontextová gramatika a zásobníkový převodník. Čtenář si jistě může sám zkontrolovat, jak souvisí přidávání dalších takových převodníků s možností podobným způsobem generovat i množiny řetězů tvaru x^n s n vyšším než 2.

$G_1 = (A, A_t, S, P)$, kde $A = \{S, B, a, b, c\}$, $A_t = \{a, b, c\}$,
P obsahuje právě tato pravidla: $S \rightarrow hBh$
$B \rightarrow aBa$
$B \rightarrow bBb$
$B \rightarrow c$

Tab. 1

1. (Karel R_{ag} (vědět$_{min}$ R'_{pat} (táta R_{ag} ((přivézt$_{následˇ}$ R'_{adres} máma) R'_{pat} dárek))))
2. (Karel$_{podmět}$ (vědět$_{min}$ (táta$_{podmět}$ ((přivézt$_{bud,předmět}$ máma$_{nepř.předm.}$) dárek$_{předmět}$))))
3. Karel$_{1.pád}$ vědět$_{3.os.min.}$ že táta$_{1.pád}$ přivézt$_{3.os.bud.}$ máma$_{3.pád}$ dárek$_{4.pád}$

Tabulka obsahuje zápisy věty „Karel věděl, že táta přiveze mámě dárek" na rovinách 1. významové, 2. povrchové stavby, 3. morfologické (vždy po průchodu dvěma zásobníkovými automaty); zápisy jsou zde v mnoha ohledech zjednodušeny; funktory tvaru R_i rozlišují vedle složkové struktury (vyznačené závorkováním) i typ syntaktické závislosti (agens, patiens, adresát, determinace adverbiální) a její směr (u funktoru bez čárky je řídícím slovem jeho pravý argument, u čárkovaného funktoru — levý).

*) V jeho rámci byla matematicky zpracována i řada jednotlivých lingvistických otázek (viz stati v Prague Bulletin of Mathem. Linguistics, vyd. dvakrát ročně MFF UK, kde se ukazuje, že i lingvista se může naučit pracovat s postupem definice — teorém — důkaz; širší dosah tu má zejm. originální zpracování sémantiky slovesného času, jazykového záporu, dále rozbor kontextového zapojení, typů doplnění slovesa, slovesné modality aj.).

Zásobníkový automat P_1 lze ve zkratce charakterizovat tabulkou, která odpovídá jeho definující funkci (s_0 je zároveň stavem výchozím a koncovým; v je za a, b):

čtený symbol na vstupu	čtený symbol v zásobníku	momentální vnitřní stav	následující vnitřní stav	symbol ukládaný do zásobníku	symbol tištěný na výstupu
h		s_0	s_1		h
v		s_1	s_1		v
c		s_1	s_2	c	
v		s_2	s_2	v	
h		s_2	s_3		
	v	s_3	s_3		v
	c	s_3	s_0		h

Vstupní řetěz považujeme za přijatý tímto automatem (tj. za prvek jeho vstupního jazyka), jestliže po přečtení řetězu automat přečte i všechny symboly, které má uloženy v zásobníku, a po přečtení posledního z nich bude ve vnitřním stavu s_0; je tedy zřejmé, že množina řetězů tvaru $hxc\hat{x}h$ je jen vlastní podmnožinou vstupního jazyka našeho automatu, který přijme i všechny ostatní řetězy tvaru $hxcyh$, kde x a y jsou řetězy nad danou abecedou. Proto také výstupní jazyk našeho automatu (který je vždy jazykem nekontextovým) není shodný s výstupním jazykem systému jako celku, tj. s množinou řetězů, na které budou automatem P_1 přeloženy výstupní řetězy gramatiky G_1. Zásobníkové automaty užívané pro generativní popis češtiny jsou podrobněji zpracovány v [13].

Třída všech nekontextových gramatik, z nichž každá je spojena s konečnou posloupností automatů uvedeného druhu, má tedy větší generativní sílu než třída všech nekontextových gramatik; zároveň však zůstává v tomto smyslu slabší než třída všech gramatik kontextových. Tento typ popisu (tzv. funkční generativní popis) ukazuje, že přirozený jazyk — pokud je takto z empirického hlediska adekvátně popsán — má důležité strukturní vlastnosti, které ho zařazují do matematicky zajímavé třídy objektů. Vlastnosti takto zjištěné se pak mohou stát východiskem pro studium řady otázek týkajících se přirozených jazyků a jejich užívání z hlediska lingvistického, psychologického, i z hlediska technických aplikací.

3. Aplikace lingvistiky související s uplatněním samočinných počítačů zdaleka nezahrnují jen strojový překlad, jehož příprava je ovšem daleko dlouhodobější záležitostí, než se zejména neodborníci v padesátých letech domnívali; o jednom z nejpokročilejších dnešních systémů (rusko-francouzském překladu skupiny z Grenoblu) pojednává práce vynikajícího matematika, prof. VAUQUOISE [14]. Hlavním cílem uvedených aplikací obecně je však něco jiného — především odstranění dosavadních obtíží, se kterými se dnes setkává uplatnění samočinných počítačů v nejrůznějších oborech a které vyžadují, aby mezi počítačem a jeho uživateli zprostředkovaly armády speciálně školených od-

borníků — nejen programátorů (o které je snad všude ve světě nouze), ale i např. indexátorů nebo anotátorů, kteří zpracovávají literaturu z řady technických a vědeckých oborů, tak aby údaje bibliografické a perspektivně i věcné byly vhodně zakódovány pro vstup do počítačových informačních systémů (kterým se pak říká automatické, ačkoli vlastní zpracování údajů na vstupu, jak řečeno, zůstává úkolem člověka).

Není divu, že se stále víc uplatňuje snaha svěřit počítači úkoly spojené s komunikací mezi ním a člověkem; sestavují se programy zajišťující, že se počítač sám orientuje v odborném textu a objevují se také první pokusy, jak nynější situaci v programování zlepšit tím, že by počítače postupně v potřebné míře zvládly jazyk lidský, místo aby se lidé stále museli pracně učit novým a novým jazykům potřebným jen k využití počítačů.*) Uplatnění angličtiny jako programovacího jazyka je podle některých [15] dnes už zcela reálným cílem výzkumu, a místo lingvistiky — nejen teoretické, jak jsme o ní výše mluvili, ale i tzv. strojové nebo počítačové lingvistiky, přímo zaměřené na nové technické možnosti — v rámci takového výzkumu, který zabezpečí další přibližování programovacích jazyků jazykům přirozeným, nebude bezvýznamné, srov. [16] a literaturu tam uvedenou.

Pokud jde o češtinu, jsme ovšem odkázáni sami na sebe, a je pochopitelné, že usilujeme také o co nejlepší teoretickou úroveň výzkumu u nás. Na základě lingvistického i formálního zpracování uvedeného v odd. 2 byla v Centru numerické matematiky MFF UK vypracována automatická analýza českých slovních tvarů, původně pro Minsk 22, nyní se programuje v jazyce PL/1; to znamená, že byly vyřešeny otázky automatického přiřazení jednotlivých slovních tvarů příslušným slovům (jako slovníkovým jednotkám), a že tedy automatické sestavování rejstříků, konkordancí a slovníků pro češtinu a příbuzné jazyky už je jen otázkou větší nebo menší pracnosti, bez zásadních problémů. Také automatická syntéza české věty už v dost rozsáhlém modelu existuje jako program pro počítače třetí generace, a stroje tedy již začínají „mluvit česky", vytvářet české věty; jde ovšem o gramatickou správnost vět, nikoli o smysluplnost, takže nepovažujeme za chybu, že stroj vedle vět plně přijatelných, jako „Paměť dovede být nějakou dobu dlouhá" nebo „Fuj!" sestavil i věty jako „Žena se dívá před leden". Strojové experimenty se syntézou vět (jaké zatím ve srovnatelném rozsahu neexistují pro jiný jazyk, než je čeština a angličtina) umožní doplnit mezery, bez počítače většinou těžko zjistitelné, a kromě toho program syntézy zahrnuje podrobně zpracované roztřídění jednotek různých rovin popisu. Je proto nyní i automatická analýza věty (nejen slovního tvaru) reálným úkolem, a můžeme doufat, že výzkum českých matematiků a lingvistů i zde bude mít své místo mezi prvními, jaké mu už po řadu let přiznávají i spolupracující odborníci v rámci plánu společného výzkumu zemí RVHP, i specialisté západní (viz např. [17]).

*) Nemůžeme se tu zabývat oboustranně plodným vztahem mezi strojovou lingvistikou a umělým intelektem, popř. robotikou; je zřejmé, že automatická analýza a syntéza vět přirozeného jazyka je i zde jedním z potřebných předpokladů.

Literatura

[1] CHOMSKY N.: *Syntactic Structures*, Haag 1957; český překl. *Syntaktické struktury*, Praha 1966.

[2] CHOMSKY N.: *On Certain Formal Properties of Grammars*. Information and Control *2* (1959), 137—157.

[3] CHOMSKY N.: *Studies on Semantics in Generative Grammar*, Haag 1972.

[4] POSTAL P. M.: *Constituent Structures*, Haag 1964; čes. překlad připojen k překladu knihy J. J. KATZE a P. M. POSTALA, *Celistvá teorie lingvistických popisů*, Praha 1967, 193—293.

[5] BAR-HILLEL Y.: *Language and Information*, Reading, Mass., 1964.

[6] GROSS M.: *On the Equivalence of Models of Language Used in the Fields of Machine Translation and Information Retrieval*. Information Storage and Retrieval *2* (1964), 43—57.

[7] MCNAUGHTON R.: *The Theory of Automata, A Survey*. Ve sb. Advances in Computers *2* (1961), New York, 379—421.

[8] SGALL P.: *Functional Sentence Perspective in a Generative Description*. Ve sb. Prague Studies in Mathematical Linguistics *2* (1967), Praha 203—225.

[9] GORALČÍKOVÁ A.: *On One Type of Dependency Grammars*. Prague Bulletin of Mathematical Linguistics *21* (1974), 11—26.

[10] *Úvod do algebraické lingvistiky* (kolektiv autorů). Skriptum MFF UK, Praha 1974.

[11] *Studie z transformační gramatiky I*, Praha 1975; *II*, Praha 1976.

[12] PLÁTEK M.: *On One System of Sets of Languages Close to Context-Free Languages*. Teorie a metoda *6* (1974), 103—120.

[13] GORALČÍKOVÁ A. a L. NEBESKÝ: *On a Possible Application of Pushdown Store Transducers*. Prague Bulletin of Mathematical Linguistics *9* (1968), 47—68; 10 (1968), 3—27.

[14] VAUQUOIS B.: *La traduction automatique à Grenoble*, Paříž 1975.

[15] SCHLESINGER J.: *English as a Programming Language*, rozmn. pro mezin. konf. o strojové lingvistice, Ottawa 1975.

[16] SGALL P. a E. HAJIČOVÁ: *A Linguistic Approach to Information Retrieval I*. Information Storage and Retrieval *10* (1974), 411—417.

[17] KLEIN W. a A. VON STECHOW: *Functional Generative Grammar in Prague*. Předmluva ke stejnojmennému sborníku, Kronberg/Taunus 1974, VII—XXX.

[18] JIRKŮ P.: *Towards an Integrated Theory of Formal and Natural Languages*. Kybernetika *11* (1975), 91—100.

[19] SGALL P.: *K některým otázkám sémantiky věty*. Slovo a slovesnost *37* (1976), 184—194.

[20] NOVOTNÝ M.: *Problémy matematické lingvistiky řešené československými matematiky*. Pokroky matematiky, fyziky a astronomie *18* (1973), 311—321.