Pokroky matematiky, fyziky a astronomie

Miroslav Novotný

Problémy matematické lingvistiky řešené československými matematiky

Pokroky matematiky, fyziky a astronomie, Vol. 18 (1973), No. 6, 311--321

Persistent URL: http://dml.cz/dmlcz/139541

Problémy matematické lingvistiky*) řešené československými matematiky

Miroslav Novotný, Brno

Všechny vědní disciplíny se dnes pokoušejí o aplikaci exaktních zvláště matematických metod. Lingvistika studuje jazyk, tedy objekt velmi složitý. Má-li být studium jazyka matematickými metodami úspěšné, je nutno od některých jeho vlastností abstrahovat, tj. přirozený jazyk nahradit jeho modelem. Takový model je tedy matematický objekt, který lze lingvisticky interpretovat.

Protože je jazyk objekt s početnými soubory různých jednotek, bylo přirozené aplikovat naň statistické metody. Byly zkoumány relativní frekvence slov daného jazyka a sestavovány frekvenční slovníky. Ty dávají důležité informace o tom, jak často se libovolné slovo daného jazyka ve větách tohoto jazyka vyskytuje; mohou tedy vydatně pomoci při sestavování učebnic cizích jazyků. Jazyk má také své specifické statistické zákonitosti. Přiřaďme ke každému slovu jazyka s jeho relativní frekvenci $f(s)$ a předpokládejme, že relativní frekvence dvou různých slov jsou různé. Slova pak lze seřadit do posloupnosti (s_i), kde $i < j$, právě když $f(s_i) > f(s_j)$. Existuje pak číslo c tak, že $f(s_i) = c/i$ pro každé přirozené i. To je tzv. zákon Zipfův; platí jen přibližně a byl v literatuře různě upravován.

Úvahy tohoto druhu vedly k vzniku tzv. *statistické* nebo *kvantitativní lingvistiky*. Při řešení jednotlivých úloh se jazyk nahrazuje statistickým modelem, tj. pravděpodobnostním polem, jež bylo zkonstruováno užitím statistických metod. Na ně se pak aplikují metody teorie pravděpodobnosti a informace.

Koncem padesátých let došlo k prvním pokusům o strojový překlad textu z jednoho jazyka do druhého. Bylo tedy v lingvistice použito počítačů. Užívání matematických strojů k řešení lingvistických problémů je obsahem tzv. *strojové lingvistiky*.

Při strojovém překladu z jednoho jazyka do druhého se v první fázi analyzuje daný text. Má-li se to provést pomocí počítače, je nutno vycházet z formálních kritérií. To vedlo k studiu tzv. analytického modelu jazyka.

Mějme množinu V; symbolem V^* označíme *volný monoid nad* V, tj. množinu všech konečných posloupností prvků množiny V včetně posloupnosti prázdné Λ s binární operací zřetězení. Jsou-li $x = (x_i)_{i=1}^m$, $y = (y_i)_{i=1}^n$ dvě konečné posloupnosti prvků z V, pak operace *zřetězení* k nim přiřazuje posloupnost $xy = (z_i)_{i=1}^{m+n}$ tak, že $z_i = x_i$ pro $i = 1, 2, \ldots, m$ a $z_i = y_{i-m}$ pro $i = m + 1, m + 2, \ldots, m + n$. Obvykle $x \in V$ identifikujeme s jednočlennou posloupností $(x) \in V^*$. Dostáváme tak $V \subseteq V^*$; zároveň nám tato identifikace spolu s asociativitou operace zřetězení umožňuje každé $x = (x_i)_{i=1}^m \in V^*$, $x_i \in V$ pro $i = 1, 2, \ldots, m$, vyjádřit ve tvaru $x = (x_1)(x_2)\ldots(x_m) = x_1x_2\ldots x_m$. Prvky množiny V^* nazýváme řetězy. Pro každé $x \in V^*$ tedy existuje $m \geqq 0$ celé a prvky

*) Podle referátu předneseného na schůzi vědeckého kolegia matematiky ČSAV dne 27. 6. 1972.

$x_1, x_2, \ldots, x_m \in V$ tak, že $x = x_1 x_2 \ldots x_m$. Klademe pak $|x| = m$; číslo m pak nazýváme *délkou řetězu* x.

Je-li V množina a $L \subseteq V^*$ libovolné, pak uspořádanou dvojici (V, L) nazýváme *analytickým modelem jazyka* nebo prostě *jazykem*. Prvky množiny V interpretujeme jako slovní tvary, prvky množiny L jako správné věty jazyka. Díváme-li se tedy na přirozený jazyk jako na analytický model jazyka, abstrahujeme od všech znalostí o tomto jazyce a ponecháváme si jen dvě základní: dovedeme rozeznat, co je a co není slovní tvar a co je a co není správná věta. Vedle toho lze předpokládat, že slovní tvary významově blízké budou zapsány podobnými symboly a že je tedy dán rozklad množiny V v třídy, jimž říkáme *tvarové soubory* nebo *paradigmata* tak, že každá třída se skládá ze slovních tvarů, které jsou si významově blízké.

Je-li např. (V, L) přirozený jazyk se skloňováním a časováním, pak takový tvarový soubor se skládá ze všech slovních tvarů, které lze z téhož základního slovního tvaru odvodit skloňováním nebo časováním. V češtině je např. takovým tvarovým souborem množina {*les, lesa, lesu, lese, lesem, lesy, lesů, lesům, lesích*}.

Tvarové soubory jsou příklady tzv. *gramatických kategorií*; jimi rozumíme třídy rozkladu množiny všech slovních tvarů, které mají gramatický význam. Další gramatické kategorie, tzv. *rodiny*, lze sestrojit užitím znalosti množiny L: Dva prvky $x, y \in V$ náleží do téže rodiny, když při libovolném $u, v \in V^*$ jsou podmínky $uxv \in L$, $uyv \in L$ spolu ekvivalentní. To značí: dva slovní tvary x, y náleží do téže rodiny, právě když lze slovní tvar x v libovolné správné větě nahradit slovním tvarem y tak, že vznikne správná věta, a naopak. V češtině např. slovní tvary *městem* a *místem* náleží do téže rodiny.

Při analýze textu mechanickým zařízením je třeba slovní tvary vhodně třídit, tj. zařazovat je do tříd jistých rozkladů. Nejhrubší třídění, jehož se užívá v tradiční gramatice, třídí slovní tvary podle druhů slov; k tomu je však třeba znát význam slovních tvarů.

Jsou tedy přirozené pokusy nahradit toto třídění slovních tvarů jiným tak, aby se slovní tvary daly třídit mechanicky podle formálních kritérií. Příklady rozkladů, při jejichž konstrukci nebylo třeba operovat s významem vět a slovních tvarů, byly rozklad v tvarové soubory a rozklad v rodiny. Sestrojíme-li nějaký rozklad na množině V tím, že vycházíme jen z těchto dvou rozkladů a ze znalosti množin V a L, je zaručeno, že výsledný rozklad byl definován bez užití významů slovních tvarů a vět. Zajímají nás především rozklady, které jsou hrubší než rozklad v tvarové soubory a také než rozklad v rodiny. Tyto otázky tedy vedou ke studiu analytického modelu jazyka (V, L), speciálně pak ke studiu konstrukcí, které k libovolnému rozkladu množiny V přiřazují rozklad hrubší. Touto problematikou se zabývali zejména O. S. KULAGINOVÁ ([23]) a S. MARCUS ([24], [25] [37],).

Vedle tohoto analytického modelu jazyka se studoval ještě tzv. generativní model jazyka: Buď U množina, buď dále $V \subseteq U$, $S \subseteq U^*$, $R \subseteq V^* \times V^*$, nechť množiny U, S, R jsou konečné. Pak se uspořádaná čtveřice $G = \langle U, V, S, R \rangle$ nazývá *gramatika*. Prvky v R nazýváme *pravidly*; je-li $(y, x) \in R$, pak y nazýváme *levou* a x *pravou stranou* tohoto *pravidla*. Prvky v množině S nazýváme *počátečními řetězy*. Gramatika G

pracuje tak, že vyjdeme z libovolného řetězu $s \in S$ a naň aplikujeme některé pravidlo $(y, x) \in R$. To znamená, že řetěz s vyjádříme jako součin $s = uyv$ tří činitelů, z nichž prostřední je roven levé straně našeho pravidla. Tento činitel pak nahradíme pravou stranou tohoto pravidla a utvoříme tedy řetěz uxv. Na tento řetěz aplikujeme podobně další pravidlo a takto pokračujeme tak dlouho, až dostaneme řetěz $z \in V^*$ složený z prvků množiny V, které nazýváme *terminálními symboly*. Označme $\mathscr{L}(G)$ množinu všech takto sestrojených řetězů; jazyk $(V, \mathscr{L}(G))$ pak nazýváme *jazykem generovaným gramatikou G. Generativním modelem jazyka* pak rozumíme jazyk generovaný nějakou gramatikou.

Tyto neformálně zavedené pojmy nyní zavedeme přesně takto: Je-li $y, x \in U^*$, pak klademe $y \underset{R}{\Rightarrow} x$, existují-li $u, v \in U^*$, $(t, z) \in R$ tak, že $y = utv$, $uzv = x$. Dále pro x, $y \in U^*$ píšeme $y \overset{*}{\underset{R}{\Rightarrow}} x$, existuje-li $n \geqq 0$ celé a řetězy $t_0, t_1, \ldots, t_n \in U^*$ tak, že $y = t_0$, $t_n = x$ a $t_{i-1} \underset{R}{\Rightarrow} t_i$ pro $i = 1, 2, \ldots, n$. Konečně klademe $\mathscr{L}(G) = \{x;\ x \in V^*$ a existuje $s \in S$ tak, že $s \overset{*}{\underset{R}{\Rightarrow}} x\}$.

Tento postup má svůj původ v tradiční gramatice. Místo symbolu **j** pro jednoduchou větu lze napsat **sp**; zde **s** je symbol pro podmět, **p** pro přísudek. Dále je možno **s** nahradit symbolem **n**; to je symbol pro podstatné jméno mužského rodu v jednotném čísle a v 1. pádě. Symbol **n** je možno nahradit řetězem **an**; zde je **a** symbol pro přídavné jméno mužského rodu v jednotném čísle a v 1. pádě. Konečně je možno symbol **p** nahradit symbolem **v** pro nepřechodné sloveso v 3. osobě jednotného čísla a v mužském rodě. Po těchto náhradách dostaneme ze symbolu **j** schéma věty. Skutečnou českou větu pak dostaneme, nahradíme-li každý ze symbolů **n**, **a**, **v** libovolným jemu odpovídajícím slovním tvarem. Tak lze např. symbol **n** nahradit slovními tvary *muž*, *chlapec*, *žák*, ...; symbol **a** slovními tvary *velký*, *malý*, *silný*, *slabý*, ...; symbol **v** slovními tvary *stojí*, *stál*, *jde*, *šel*, *sedí*, *seděl*, ... Každý z těchto slovních tvarů ovšem považujeme za nedělitelný symbol.

Označme nyní V množinu všech slovních tvarů, které se zde vyskytly, položme dále $U = V \cup \{\mathbf{j}, \mathbf{s}, \mathbf{p}, \mathbf{n}, \mathbf{a}, \mathbf{v}\}$, $S = \{\mathbf{j}\}$, $R = \{(\mathbf{j}, \mathbf{sp}), (\mathbf{s}, \mathbf{n}), (\mathbf{n}, \mathbf{an}), (\mathbf{p}, \mathbf{v}), (\mathbf{n}, \textit{muž}), (\mathbf{n}, \textit{chlapec})$ $(\mathbf{n}, \textit{žák}), \ldots, (\mathbf{a}, \textit{velký}), (\mathbf{a}, \textit{malý}), (\mathbf{a}, \textit{silný}), (\mathbf{a}, \textit{slabý}), \ldots, (\mathbf{v}, \textit{stojí}), (\mathbf{v}, \textit{stál}), (\mathbf{v}, \textit{jde}),$ $(\mathbf{v}, \textit{šel}), (\mathbf{v}, \textit{sedí}), (\mathbf{v}, \textit{seděl}), \ldots\}$. Položme $G = \langle U, V, S, R\rangle$. Pak G je gramatika, která generuje část českého jazyka. Vidíme, že první čtyři pravidla zachycují jisté velmi obecné zákonitosti českého jazyka: 1. Jednoduchá věta se skládá z podmětu a z přísudku. 2. Podmětem může být podstatné jméno mužského rodu v jednotném čísle a v 1. pádě. 3. Před každé podstatné jméno je dovoleno položit přídavné jméno, které se s podstatným shoduje v čísle, v rodě a v pádě. 4. Přísudkem může být nepřechodné sloveso v třetí osobě jednotného čísla, bylo-li podmětem podstatné jméno v jednotném čísle. Tato pravidla nevedou ještě k skutečné české větě, nýbrž jen k jejímu obecnému schématu, tj. k posloupnosti utvořené z gramatických kategorií **n**, **a**, **v**. Skutečná věta se dostane teprve tehdy, když každou z těchto gramatických kategorií nahradíme libovolným slovním tvarem, který do ní náleží.

Ukážeme to na konkrétním příkladě:

j $\underset{R}{\Rightarrow}$ **sp** $\underset{R}{\Rightarrow}$ **np** $\underset{R}{\Rightarrow}$ **nv** $\underset{R}{\Rightarrow}$ **anv** $\underset{R}{\Rightarrow}$ **aanv** $\underset{R}{\Rightarrow}$ *malý* **anv** $\underset{R}{\Rightarrow}$ *malý* **a** *chlapec* **v** $\underset{R}{\Rightarrow}$ *malý slabý chlapec* **v** $\underset{R}{\Rightarrow}$ *malý slabý chlapec seděl*

Buď $G = \langle U, V, S, R\rangle$ gramatika. Jestliže $y \in (U - V)^* - \{\Lambda\}$ pro každé pravidlo $(y, x) \in R$ a jestliže existuje $s \in U - V$ tak, že $S = \{s\}$, pak říkáme, že G je *frázová strukturní gramatika* nebo též *gramatika typu* 0. Je-li $|y| \leqq |x|$ pro každé pravidlo $(y, x) \in R$ frázové strukturní gramatiky $G = \langle U, V, \{s\}, R\rangle$, pak říkáme, že G je *kontextová gramatika* nebo též *gramatika typu* 1. Je-li $y \in U - V$ pro každé pravidlo $(y, x) \in R$ frázové strukturní gramatiky $G = \langle U, V, \{s\}, R\rangle$, pak říkáme, že G je *nekontextová gramatika* nebo též *gramatika typu* 2. Buď $G = \langle U, V, \{s\}, R\rangle$ frázová strukturní gramatika taková, že pro každé pravidlo $(y, x) \in R$ jsou splněny tyto podmínky: (a) $y \in U - V$; (b) $x = \Lambda$ nebo $x \in V$ nebo $x = uv$, kde $u \in U - V$, $v \in V$; pak G nazýváme *regulární gramatikou* nebo též *gramatikou typu* 3. Jazyk (V, L) nazýváme *jazykem typu* i $(i = 0, 1, 2, 3)$, je-li generován gramatikou typu i. Jazyky typu 0 též nazýváme *konstruktivními*, jazyky typu 1 *kontextovými*, jazyky typu 2 *nekontextovými* a jazyky typu 3 *regulárními*. Pojem gramatiky, klasifikace gramatik a klasifikace jazyků pochází od N. CHOMSKÉHO ([2], [3]).

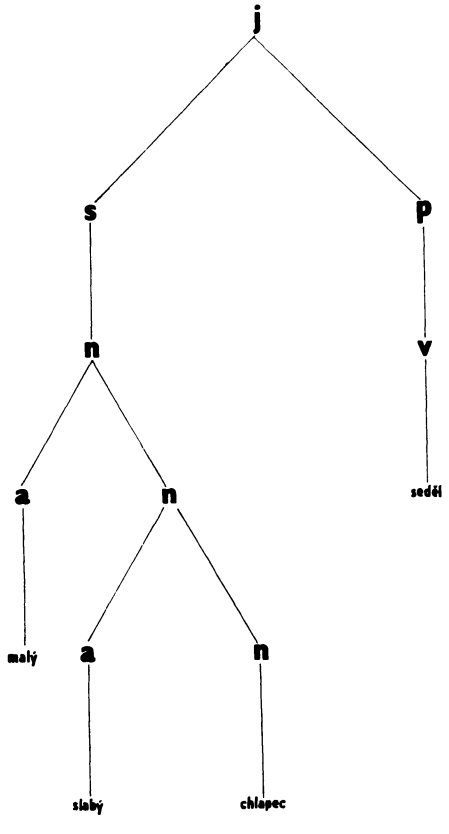
Generování věty nekontextovou gramatikou, jejíž pravidla nepřipouštějí prázdný řetěz na pravých stranách, lze zachytit přirozeným způsobem graficky: Vyjdeme z počátečního symbolu a přiřadíme mu bod ve svislé rovině. Při prvním kroku generování se tento počáteční symbol nahradil neprázdným řetězem. K jeho symbolům přiřadíme body v naší rovině, které umístíme v přirozeném pořádku pod bod odpovídající počátečnímu symbolu a spojíme hranami s bodem, který byl přiřazen k počátečnímu symbolu. Podobně postupujeme dále. Vzniklý graf, jehož uzly jsou popsány symboly naší gramatiky, nazýváme *frázový ukazatel* dané *věty*. Protože gramatika z našeho příkladu je nekontextová a její pravidla nepřipouštějí na pravých stranách prázdný řetěz, můžeme konstrukci frázového ukazatele ilustrovat na větě *malý slabý chlapec seděl* (viz obr. 1). Náleží-li ke každé větě generované danou nekontextovou gramatikou, jejíž pravidla nepřipouštějí prázdný řetěz na pravé straně, právě jeden frázový ukazatel, říkáme, že gramatika je *jednoznačná*; v opačném případě ji nazýváme *víceznačnou*.

Problematice gramatik a jejich jazyků je věnována rozsáhlá literatura [12], [19].

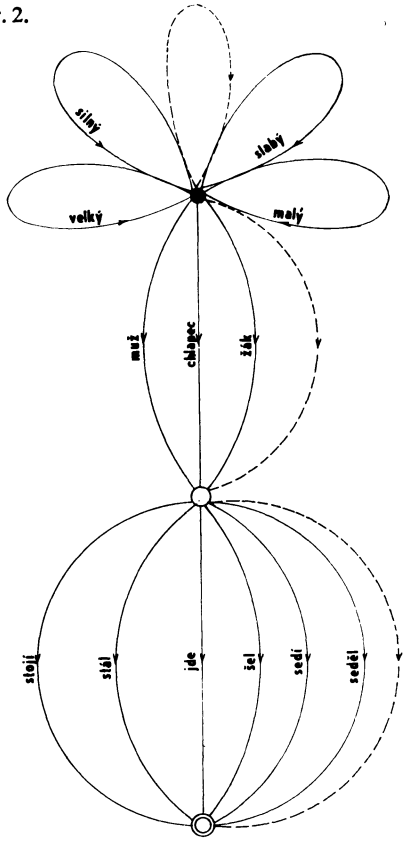
Gramatika jako zařízení produkující věty jazyka imituje činnost mluvící osoby. Osoba poslouchající a rozlišující správné věty jazyka od ostatních posloupností slovních tvarů se imituje tzv. akceptorem. Připomeňme zde pojem konečného nondeterministického akceptoru: Jsou dány dvě konečné množiny, a to množina S, jejíž prvky nazýváme *stavy*, a množina V, jejíž prvky odpovídají slovním tvarům. V množině S je vyznačena množina I *počátečních stavů* a množina F *koncových stavů*. Konečně je dáno zobrazení f množiny $S \times V$ do množiny 2^S všech podmnožin množiny S. Uspořádanou pětici $N = \langle S, V, f, I, F\rangle$ nazýváme pak *konečným nondeterministickým akceptorem*. Takový akceptor lze vyjádřit graficky tak, že vyznačíme stavy jako body v rovině, dále vyznačíme množinu všech počátečních stavů a množinu všech koncových stavů. Ke každému stavu $s \in S$ a ke každému symbolu $a \in V$ sestrojíme množinu orientovaných hran vycházejících z bodu s jdoucích do všech bodů množiny $f(s, a)$ a popsaných symbolem a. Řetěz $x \in V^*$ je *akceptorem N přijat*, existuje-li v našem grafickém znázornění sled vycházející

z některého počátečního stavu a vcházející do některého koncového stavu tak, že popis hran tohoto sledu v příslušném pořádku dává řetěz x. Označme symbolem $\mathscr{L}(N)$ množinu všech řetězů přijatých akceptorem $N = \langle S, V, f, I, F \rangle$. Pak jazyk $(V, \mathscr{L}(N))$ nazýváme *jazyk přijatý akceptorem* N.

Obr. 1.



Obr. 2.



Jako příklad uveďme akceptor znázorněný na obr. 2. Tento akceptor má celkem tři stavy; počáteční je vyznačen plným kroužkem, koncový stav je vyznačen kroužkem dvojitým, zbývající stav kroužkem jednoduchým. Tento akceptor přijímá právě všechny věty generované gramatikou popsanou na začátku tohoto článku. Lze ukázat, že třída všech jazyků přijatých konečnými nondeterministickými akceptory je rovna třídě všech regulárních jazyků ([36], [12], [19]).

Všechny tyto úvahy mají společné jádro: Objekty zde studované (jazyk, gramatika, akceptor) lze chápat jako obecné algebraické struktury. Proto úvahy tohoto druhu zařazujeme do tzv. *algebraické lingvistiky.*

Matematickou lingvistiku můžeme charakterizovat jako vědní disciplínu, v níž se lingvistické problémy řeší užitím matematických metod. Podle povahy těchto metod ji tedy můžeme rozdělit na lingvistiku statistickou neboli kvantitativní, na lingvistiku strojovou a na lingvistiku algebraickou. Tím je obsah těchto disciplín dán jen rámcově. Zatímco lingvisté zahrnují do matematické lingvistiky jakékoliv užití matematických metod v lingvistice (tímto hlediskem se řídí např. Referativnyj žurnal matematika), matematikové obvykle chápou tuto vědní disciplínu úžeji a rozumějí matematickou lingvistikou souhrn takových matematických výsledků, které řeší lingvistickou problematiku a zároveň obohacují matematické teorie (tímto hlediskem, které vylučuje z matematické lingvistiky užití standardních matematických metod při řešení lingvistických problémů, se řídí např. časopis Mathematical Reviews). Chceme-li považovat matematickou lingvistiku za matematickou vědní disciplínu, musíme se postavit na toto druhé stanovisko. Odtud ovšem plynou jisté potíže s hodnocením výsledků: Hodnocení z lingvistického hlediska se může lišit od hodnocení podle měřítek matematických. Matematické výsledky bývají obecného charakteru a pro konkrétní jazyky obvykle dávají jen málo; výsledky, jež považují lingvisté za důležité, se matematikům jeví většinou jen jako snůška fakt, případně definic.

Podáme nyní přehled stavu matematické lingvistiky v ČSSR jako matematické disciplíny. Budeme si proto všímat především prací, které obohacují matematické teorie; jen velmi stručně se dotkneme užití standardních matematických metod v lingvistice. Není možno v tomto článku, který má informativní charakter, popsat veškeré matematické výsledky, jichž se v matematické lingvistice u nás dosáhlo. K tomu by bylo potřeba mnohem mohutnějšího pojmového aparátu, než je aparát, který jsme právě zavedli. Proto si problematiku shrneme do několika charakteristických tematických okruhů a v každém z nich vyložíme jen základní výsledky, jichž se zde dosáhlo. Tento průřez prací československých matematiků byl pořízen hlavně podle referátů v časopise Mathematical Reviews; proto se týká převážně výsledků starých dva roky a starších. Autor se domnívá, že i přes tuto neúplnost může být zdrojem informací o problematice matematické lingvistiky pěstované v Československu.

Statistická vyšetřování jazyka prováděli téměř výhradně lingvisté. Všímali si většinou frekvence různých jazykových jevů, pokoušeli se najít statistická kritéria pro klasifikaci textů apod. Aplikovali přitom vesměs standardní matematické metody ([1]).

Podobně byla i strojová lingvistika doménou lingvistů. Jejich úsilí směřovalo k sestavení programu pro strojový překlad z angličtiny do češtiny. Takový program byl sestaven kolektivem pracovníků filosofické fakulty Karlovy university a pokusně byly podle něho texty překládány ([20]). Je do jisté míry problematické, zda lze sestavení programu považovat za výkon obohacující matematické teorie. Z matematického hlediska by bylo žádoucí, aby po tomto programu následoval důkaz věty, že užitím tohoto programu je možno každou správnou anglickou větu přeložit v správnou českou větu téhož významu. Jenomže matematické definice pojmů „správná anglická věta", „správná česká věta", „význam správné anglické věty" a „význam správné české věty" nemáme. Proto je nutno se spokojit se sestavením tohoto programu a na příkladech ověřovat, že k správným anglickým větám přiřazuje správné české věty téhož významu. Tato obtíž se neomezuje jen na překlady z jednoho přirozeného jazyka do druhého, nýbrž

objevuje se také při překladech z jednoho programovacího jazyka do druhého a velmi ztěžuje hodnocení takových programů z matematického hlediska.

Na rozhraní strojové a algebraické lingvistiky je zajímavý pokus K. PALY ([35]), který spočíval v tom, že na základě analýzy jistých českých textů byla sestavena gramatika generující české věty. K této gramatice byl sestrojen program, který umožňoval generovat české věty na samočinném počítači. Tento program dává poměrně dobré výsledky; většina vět, které se takto dostanou, je správná a má smysl.

Otázkami, o kterých jsme se právě zmínili, jsou motivovány snahy lingvistů v algebraické lingvistice. Podle představ P. SGALLA si lze jazyk představit jako útvar o několika rovinách. V nejvyšší rovině se užitím gramatiky dostanou obecná schémata vět, v nichž se jednotlivé slovní tvary objevují jen v základní podobě, avšak jsou doprovázeny dalšími pomocnými symboly, jež umožňují vyjádřit smysl věty. Z vyšší roviny se na nižší dostaneme transformací všech vět vyšší roviny; tato transformace se provádí jistými automaty. V nejnižší rovině dostaneme větu v její fonetické podobě. P. Sgall zejména popsal gramatiku i automaty, které se v této teorii vyskytují ([38]).

Nyní přistoupíme k přehledu činnosti matematiků v algebraické lingvistice.

Řada prací byla věnována studiu analytického modelu jazyka. Tyto práce navazují zejména na práce Kulaginové a Marcusovy, v nichž autoři konstruují gramatické kategorie. V rámci své teorie Marcus formuloval řadu problémů, z nichž některé řešil B. ZELINKA ([39], [40]). Kulaginová a Marcus budují teorii analytického modelu jazyka na pojmech teorie množin; tento model je proto někdy označován v lingvistické literatuře jako množinový model jazyka. M. NOVOTNÝ ([31], [32]) ukázal, že tento množinový model jazyka je algebraická struktura, která má svou algebraickou teorii, v níž lze hlavní věty Kulaginové a Marcusovy odvodit. Množinovými metodami zkoumal L. NEBESKÝ ([28]) pojem kontextu.

V souhlase se světovým vývojem studovali českoslovenští matematikové intenzívně nekontextové gramatiky.

Podmínky nutné a dostatečné pro jednoznačnost nekontextových gramatik formuloval V. FABIAN ([9]) pomocí pojmu gramatického elementu, který sám zavedl. Další pojmy, pomocí nichž bylo možno formulovat podmínky jednoznačnosti nekontextových gramatik, definoval J. GRUSKA ([14], [15]). Ten také zavedl různá kritéria pro klasifikaci nekontextových gramatik ([16]). Studoval také jisté podsystémy množin pravidel v nekontextových gramatikách; tyto podsystémy nazývá gramatickými rovinami. Ukázal např., že pro každé přirozené n existuje nekontextový jazyk tak, že každá jeho jednoznačná gramatika má více než n rovin ([17]).

Je známo, že třída nekontextových jazyků je příliš úzká pro potřeby přirozených jazyků; existují totiž přirozené jazyky, které nejsou nekontextové. Na druhé straně však nekontextové gramatiky umožňují konstrukci frázového ukazatele, který se při obecnějších gramatikách už jednoduše definovat nedá. Proto vznikly pokusy užít nekontextových gramatik ke generování jazyků, které nejsou nekontextové. Tak např. I. FRIŠ ([11]) předpokládal, že množina pravidel nekontextové gramatiky je částečně uspořádána a že se tento pořádek při užívání pravidel respektuje. Tímto způsobem dostává bohatší třídu než je třída nekontextových jazyků. E. NAVRÁTIL zase přiřadil ke každému pravidlu dané nekontextové gramatiky regulární jazyk; pravidla je dovoleno

použít jen na řetěz, který náleží k tomuto jazyku. Při tomto použití dávají nekontextové gramatiky právě všechny konstruktivní jazyky ([27]). Na druhé straně ukázal I. HAVEL ([18]), že kontextové gramatiky s jistými speciálními typy pravidel dávají jen nekontextové jazyky.

Jiný způsob, jak rozšířit generativní sílu nekontextových gramatik, je gramatická transformace. Tento pojem má svůj původ v tradiční gramatice: zde se ze základních, aktivních vět jazyka dostanou nové, pasívní jistými "transformacemi". Do algebraické lingvistiky byl pojem transformace zaveden N. CHOMSKÝM ([2]) a K. ČULÍK jej formalizoval ([6]). Jsou-li dány dvě nekontextové gramatiky, jejichž pravidla nepřipouštějí prázdný řetěz na pravých stranách, pak transformaci definuje pomocí zobrazení, jež ke každému pravidlu první gramatiky přiřazuje pravidlo druhé gramatiky. Na toto zobrazení klade jisté podmínky, které zaručují, že lze užitím tohoto zobrazení ke každému frázovému ukazateli věty generované první gramatikou přiřadit frázový ukazatel věty generované druhou gramatikou, a tedy k dané větě větu transformovanou. Podobně formalizoval Čulík pojem překladu jazyka ([7]): k zobrazení množiny pravidel gramatiky prvního jazyka do množiny pravidel gramatiky druhého jazyka přistupuje ještě zobrazení množiny symbolů gramatiky prvního jazyka do množiny symbolů gramatiky druhého jazyka. Zde se ovšem předpokládá, že k řetězům, které lze v obou gramatikách odvodit, je přiřazen jejich obraz v jisté abstraktní množině, tzv. význam. Formulují se pak pro obě zobrazení podmínky, při nichž lze přirozenou konstrukcí k frázovému ukazateli věty generované první gramatikou přiřadit frázový ukazatel věty generované druhou gramatikou tak, že obě mají týž význam. Na tyto Čulíkovy výsledky navázal J. KOPŘIVA ([21]), který zobecnil Čulíkův pojem přeložitelnosti.

Algebraická lingvistika se pokouší najít i jiné třídy jazyků, než jsou třídy 0, 1, 2, 3 Chomského hierarchie. To vede k pokusům definovat ještě jiné gramatiky, než které zavedl Chomsky. Gramatiky, v nichž místo pravidel vystupují multipravidla, tj. konečné posloupnosti obyčejných pravidel, vyšetřovali J. KRÁL ([22]) a K. ČULÍK II. ([8]). I. Friš ([10]) definoval třídu tzv. gramatizovatelných jazyků, z nichž se všechny konstruktivní jazyky dostanou jednoduchou konstrukcí.

Jak jsme již viděli, při popisu lingvistických faktů se často s výhodou používá grafů a jejich různých modifikací. Tak např. popsal K. Čulík ([4]) konečné akceptory jako pojmenované multigrafy a grafů užil při řešení problému překladu slov [5]. Zde uvažuje o množině L, na níž je dána binární relace ϱ a rozklad $\bar{L}$ tak, že ze vztahu $(x, y) \in \varrho$ plyne, že x a y náleží do různých tříd rozkladu $\bar{L}$. Třídy rozkladu $\bar{L}$ Čulík interpretuje jako slovníky různých jazyků a relaci ϱ jako relaci přeložitelnosti. Klade se otázka, zda lze konstruovat další slovník, přes který by se dalo překládat. Tato úloha se řeší zavedením pojmu sémantiky: tou se rozumí množinová reprezentace relace ϱ, tj. systém množin $\mathcal{M}$ pokrývající množinu M a zobrazení φ množiny L do $\mathcal{M}$ takové, že $(x, y) \in \varrho$, právě když $\varphi(x) \cap \varphi(y) \neq \emptyset$. Je-li dána taková sémantika, pak převodní jazyk má slovník M a relace přeložitelnosti ϱ_i ze slovníku $L_i \in \bar{L}$ do M se definuje takto: $(x, y) \in \varrho_i$, když $x \in L_i$, $y \in M$ a $y \in \varphi(x)$. Při popisu závislosti mezi slovními tvary ve větách jazyků se uplatňují tzv. závislostní stromy a mezi nimi pak zejména tzv. projektivní závislostní stromy. L. Nebeský ([30]) našel podmínky nutné a dostatečné k tomu, aby

závislostní strom byl projektivní; pokusil se také o systematické zpracování algebraických vlastností stromů ([29]).

Lze ukázat, že ne každý jazyk lze generovat gramatikou; jazyky konstruktivní, které jsou generovány gramatikami, tvoří mezi ostatními jazyky výjimky. Vzniká pak otázka, zda je možno vymezit nějakou třídu jazyků, které jsou generovány gramatikami, a popsat pak konstrukci, která ke každému jazyku této třídy přiřazuje gramatiku generující tento jazyk. Jako první řešil problém tohoto druhu A. V. GLADKIJ ([13]) pomocí tzv. konfigurací jazyků. Jeho postup zobecnil M. Novotný ([34]) takto: Buď (V, L) jazyk. Množina $R \subseteq V^* \times V^*$ se nazývá dostatečná pro (V, L), jestliže $u, v \in V^*$, $(y, x) \in R$, $uyv \in L$ implikuje $uxv \in L$. Je-li R dostatečná pro (V, L), pak pro $z \in L$ klademe $z \in B_R(V, L)$, jestliže $t \in L$, $t \overset{*}{\underset{R}{\Rightarrow}} z$ implikuje $|t| \geqq |z|$. Jazyk (V, L) se pak nazývá omezený, je-li množina V konečná a existuje-li konečná množina R dostatečná pro (V, L) tak, že množina $B_R(V, L)$ je konečná. Lze pak ukázat, že třída všech konstruktivních jazyků je tvořena právě všemi jazyky tvaru $(U \cap V, U^* \cap L)$, kde U je konečná množina a jazyk (V, L) je omezený; $\langle U \cup V, U \cap V, B_R(V, L), R\rangle$ je pak gramatika jazyka $(U \cap V, U^* \cap L)$. Tím je náš problém řešen pro třídu všech konstruktivních jazyků; podobně lze podat konstrukce gramatik pro všechny třídy Chomského hierarchie ([33]). Užijeme-li terminologie běžné mezi lingvisty, můžeme říci, že tyto práce vyšetřují vztahy mezi analytickými a generativními modely jazyka.

Chceme-li zhodnotit přínos československých matematiků k matematické lingvistice, stačí nám omezit se na lingvistiku algebraickou, protože jen v této části matematické lingvistiky naši matematikové systematicky pracují. Ve srovnání se světovými proudy v tomto oboru pozorujeme u nás poměrně veliký zájem o analytické modely jazyka; S. Marcus ([26]) vidí hlavní důvod v tom, že čeština je jazyk s bohatou flexí a že pro takové jazyky dává teorie analytických modelů netriviální výsledky. Práce československých matematiků pomáhaly dotvářet teorii analytických i generativních modelů jazyka. Tyto práce mají všechny rysy prací v tvořící se matematické disciplíně. Především lze pozorovat, že pojmy, jichž se v těchto pracích užívá, nejsou vždy ustáleny; jejich vhodnost se teprve ověřuje a ony samy se přizpůsobují jednotlivým problémům. O mnohých z těchto pojmů lze po matematické stránce dokázat jen málo. Dále najdeme v těchto pracích speciální pojmy a důkazy speciálních vět tam, kde je možno sáhnout do zásobárny teorie obecných algebraických struktur pro pojmy a věty obecnější. Konečně je třeba říci, že systematické zpracování řešených problémů nám dosud schází.

Všechny matematické disciplíny mají tendenci vyvíjet se jako samostatné logické celky. Lze tedy očekávat, že i matematická lingvistika v ČSSR bude pokračovat ve svém vývoji, a to především tam, kde se již dosáhlo podstatných výsledků, tedy v lingvistice algebraické zejména ve všech dosud pěstovaných směrech. Je žádoucí, aby ve svých studiích pokračovali jak matematikové, tak i lingvisté. Další rozvoj algebraické lingvistiky může obohatit matematiku o studium nových struktur a lingvistiku o nové metody. Bylo by vhodné, aby pěstitelé algebraické lingvistiky sledovali pokroky těch matematických disciplín, s nimiž má algebraická lingvistika styčné body, tj. hlavně teorii algebraických struktur, teorii automatů, teorii algoritmů, teorii rekursivních funkcí a teorii programovacích jazyků.

Zajímavé výsledky československých matematiků v algebraické lingvistice upoutávají zájem našich mladých matematiků. Těm však schází matematickým stylem napsaná učebnice algebraické lingvistiky, která by je rychle a bezpečně dovedla k aktuálním problémům této vědní disciplíny. Zdá se však, že hodnotných výsledků z tohoto vědního oboru je v československé matematické literatuře tolik, že lze v nejbližší době počítat s jejich zpracováním ve formě monografií.

Literatura

[1] Benešová, E., *O matematické lingvistice.* Pokroky matematiky, fyziky a astronomie *9* (1964), 335—343.

[2] Chomsky, N., *Three models for the description of language.* IRE Transactions of Information Theory V. IT — 2, No 3 (1956), 113—124.

[3] Chomsky, N., *On certain formal properties of grammars.* Information and Control *2* (1959), 137—167.

[4] Čulík, K., *Some notes on finite state languages and events represented by finite automata using labelled graphs.* Čas. pěst. mat. *86* (1961), 43—55.

[5] Čulík, K., *Ispolzovanie abstraktnoj semantiki i teorii grafov v mnogojazyčnych perevodnych slovarach.* Probl. kib. *13* (1965), 221—232.

[6] Čulík, K., *On some transformations in context-sensitive grammars and languages.* Czech. Math. J. *12* (1967), 278—311.

[7] Čulík, K., *Chorošo perevodimyje jazyki i jazyki tipa ALGOL.* Naučno techn. inf. 1967 ser. 2 No 3, 21—23.

[8] Čulík II., K., *n-ary grammars and the description of mapping of languages.* Kybernetika (Prague) *6* (1970), 99—117.

[9] Fabian, V., *Structural unambiguity of formal languages.* Czech. Math. J. *14* (1964), 394—430.

[10] Friš, I., *On Stop-Conditions in the Definitions of Constructive Languages.* Zeitschr. f. math. Logik und Grundlagen d. Math. *11* (1965), 61—73.

[11] Friš, I., *Grammars and partial orderings of the rules.* Information and Control *12* (1968), 415—425.

[12] Ginsburg, S., *The Mathematical Theory of Context Free Languages.* McGraw-Hill Book Company, New York and London, 1966.

[13] Gladkij, A. V., Konfiguracionnye charakteristiki jazykov. Probl. kib. *10* (1963), 251—260.

[14] Gruska, J., *On structural unambiguity of formal languages.* Czech. Math. J. *15* (90), (1965), 283—294.

[15] Gruska, J., *Isolable and weakly isolable sets.* Czech. Math. J. *16* (*91*), (1966), 76—90.

[16] Gruska, J., *On a classification of CF languages.* Kybernetika (Prague) *3* (1967), 22—29.

[17] Gruska, J., *Complexity and Unambiguity of Context-Free Grammars and Languages.* Information and Control *18* (1971), 502—519.

[18] Havel, I., *A note on one-sided context-sensitive grammars.* Kybernetika (Prague) *5* (1969), 186—189.

[19] Hopcroft, J. E. - Ullman, J. D., *Formal Languages and their Relation to Automata.* Addison-Wesley Publishing Co., Reading, 1969.

[20] Konečná, D. - Novák, P. - Sgall, P., *Machine Translation in Prague.* Prague Studies in Mathematical Linguistics *1* (1966), Academia, Prague, 185—193.

[21] Kopřiva, J., *Generalization of well transformation of formal languages.* Kybernetika (Prague) *2* (1966), 305—313.

[22] Král, J., *On multiple grammars.* Kybernetika (Prague) *5* (1969), 60—85.

[23] Kulagina, O. S., *Ob odnom sposobe opredelenija grammatičeskich ponjatij na baze teorii množestv.* Probl. kib. *1* (1958), 203—214.

[24] MARCUS, S., *Asupra unui model logic al părţii de vorbire.* Studii şi cercetari matematice *13* (1962), 37—62.
[25] MARCUS, S., *Algebraic Linguistics; Analytical Models.* Academic Press, New York and London, 1967.
[26] MARCUS, S., *Analytique et génératif dans la linguistique algébrique.* To Honor Roman Jacobson. Essays on the occassion of his seantieth birthday. Mouton. The Hague — Paris 1967, 1252—1260.
[27] NAVRÁTIL, E., *Context-free grammars with regular conditions.* Kybernetika (Prague) *6* (1970), 118—126.
[28] NEBESKÝ, L., *K ponjatiju kontexta.* Prague Studies in Math. Linguistics *2* (1967), Academia Prague, 179—185.
[29] NEBESKÝ, L., *Algebraic properties of trees.* Acta Universitatis Carolinae — Philologica Monographia, XXV, 1969.
[30] NEBESKÝ, L., *Projectivity in Linguistics and Planarity in Graph Theory.* Prague Studies in Math. Linguistics *5* (v tisku).
[31] NOVOTNÝ, M., *Ob algebraizacii teoretiko-množestvennoj modeli jazyka.* Probl. kib. *15* (1965), 236—244.
[32] NOVOTNÝ, M., *Algebraic Structures of Mathematical Linguistics.* Bull. Math. de la Soc. Sci. Math. de la R. S. de Roumanie *12 (60)* (1968), 87—101.
[33] NOVOTNÝ, M., *Complete Characterization of Classes of Chomsky by means of Configurations.* Acta F. R. N. Univ. Comen. — Mathematica — Mimoriadne číslo — (1971), 63—71.
[34] NOVOTNÝ, M., *O matematických modelech jazyka.* Jazykovědné aktuality. Inf. zpravodaj československ. jazykovědců *9* (1972), č. 4, 5—7.
[35] PALA, K., *Náhodné generování českých vět.* Slovo a slovesnost *29* (1968), 45—56.
[36] RABIN, M. O. - SCOTT, D., *Finite automata and their decision problems.* I. B. M. J. Res. Dev. *3* (1959), 114—125.
[37] REVZIN, I. I., *Modeli jazyka.* Izdat. Akad. Nauk SSSR, Moskva 1962.
[38] SGALL, P., *Generativní popis jazyka a česká deklinace.* Academia, Praha 1967.
[39] ZELINKA, B., *Un language adéquat non homogène dont les classes sont disjoint deux à deux.* Rev. Roum. math. pures et appl. *10* (1965), 1255—1257.
[40] ZELINKA, B., *Un language héréditaire qui n'est pas à un nombre d'états finis.* Rev. Roum. math. pures et appl. *12* (1967), 1107—1108.

Josiah Willard Gibbs

Mezi „otce" statistické mechaniky a osobnosti, které se zasloužily o termodynamiku, je jistě právem počítán J. W. Gibbs (vysl. gybs), jehož jméno nese řada pojmů a relací v těchto oborech. Uveďme si nejznámější: Gibbsův soubor a Gibbsovo rozdělení, Gibbsovo pravidlo termodynamických fází, Gibbsův paradox, Gibbsova-Helmholtzova rovnice, Gibbsův termodynamický potenciál (někdy nazývaný volnou enthalpií), Gibbsovy trojrozměrné diagramy, Gibbsova teorie fluktuací — což jistě není výčet úplný. Snad stojí ještě za zmínku, že Gibbs zavedl označení $\boldsymbol{AB}$ pro skalární a $\boldsymbol{A} \times \boldsymbol{B}$ pro vektorový součin dvou vektorů, jak je např. uvedeno v Sommerfeldově *Mechanice.* A nazve-li Sommerfeld někoho velkým termodynamikem, pak určitě toto epiteton ornans sedí.

Gibbs se narodil 11. února 1839 v New Haven v USA, absolvoval universitu v Yale r. 1858; od r. 1863 do r. 1869 pobýval studijně v Evropě (Paříž, Berlín, Heidelberg), aby dva roky po návratu do New Haven, tedy r. 1871, byl jmenován profesorem matematické fyziky na universitě v Yale. Toto místo zastával až do svého skonu 28. 4. 1903. Poslední Gibbsovo dílo *Elementary Principles in Statistical Mechanics* (1902) patří ke klasickým dílům teoretické fyzikální literatury.

Závěrem si dovoluji uvést svůj čistě privátní pohled na Gibbsovo dílo: V rámci „evropského nacionalismu" se někdy tvrdívá, že všechny velké fyzikální myšlenky přišly ze „starého kontinentu". Myslím, že nejméně dvěma Gibbsovým originálním myšlenkám bylo by možno připsat evropský charakter.

Miroslav Brdička