

údaj ukazuje, že svou větnou stavbou má vojenský jazyk extrémní postavení mezi jinými styly psané češtiny.

V této oblasti zjišťujeme též **rozložení syntaktických konstrukcí**, tj. četnosti a poměry, např. počet jednoduchých vět (počet souvětí, hlavních vět), vedlejších vět, četnosti vět oznamovacích, rozkazovacích atd. Ve výzkumu syntaktických konstrukcí je ve statistické lingvistice zatím málo zkušeností a proto můžeme očekávat, že právě zde bude nutné definovat nové stylové ukazatele.

Tento náčrt programu pro statistický výzkum vojenského jazyka je předběžný. Program vědeckého výzkumu lze plánovat pouze v nejjzákladnějších rysech. Sám výzkum přináší ve svém průběhu stále nové, zpočátku netušené problémy.

Podobně je tomu s praktickým využitím získaných poznatků. I zde můžeme pouze naznačit některé možnosti. Upozor-

ňuji zejména na úkol optimalizace, resp. **minimalizace** vojenského jazyka. Minimalizací vojenského jazyka rozumíme snížení repertoáru jazykových prostředků na nejmenší počet, který dovoluje plnit základní komunikační úlohy. Příkladem minimalizovaných jazyků mohou být mezinárodní komerční a letecké kódy, nebo tzv. Basic English.

Minimalizace vojenského jazyka by značně zjednodušila technické problémy automatizace. Nezbytným předpokladem minimalizace jsou ovšem údaje o užívaných jazykových prostředcích a jejich četnostech. Na jejich základě je možno stanovit jednotky dubletní a jednotky fidké a zakázat je používat v minimalizovaném jazyku. Vojenský jazyk, který můžeme snadno vědomě regulovat, má nejlepší předpoklady k tomu, aby byl minimalizován, popř. jinak optimálně přizpůsoben svým funkcím.

Strojový překlad a algebraická lingvistika

Doc. Dr. P. Sgall, CSc.; promovaný filolog P. Pitha,
Matematicko-fyzikální fakulta KU



Strojový překlad můžeme do určité míry chápat jako model činnosti překladatele. V prvním stadiu tohoto procesu jsou ve stroji identifikovány jednotlivé složky textu originálu (vstupního jazyka). Tomuto stadiu říkáme analýza a je možno je přirovnat k pochopení textu člověkem. Druhá část překladu, syntéza výstupního jazyka, spočívá ve výběru a uspořádání prvků výstupního jazyka tak, aby byl zachycen („pochopený“) význam analyzovaného textu.

Práce strojem je ovšem přes tuto analogii od práce lidského překladatele velmi odlišná. U stroje chybí důležitý prvek lidské činnosti, totiž předchozí zkušenost, a to jak věcná, daná všeobecným nebo i odborným vzděláním, tak jazyková. Stroj nemůže vyřešit žádnou situaci, která při přípravě SP nebyla plánována a do stroje jako potřebné údaje vložena; nemůže např. pochopit text porušený, nemůže hledat věcné poučení tam, kde by toho bylo pro

správný překlad zapotřebí.¹⁾ Jinou odlišností je nedostatek lidské paměti, která lidskému překladateli umožňuje řešit opakující se úlohy přímo, zatímco stroj musí vždy projít všemi body zdoluhavé analýzy znovu. Tento rozdíl však nelze chápat jako nevýhodu strojů, uvažujeme-li nesmírné množství operací, které je stroj schopen uskutečnit v minimální krátké době.

Stroj nepracuje s jazykem v té podobě, na kterou jsme zvyklí. Text se uvádí do stroje v podobě kódu (který může být velmi různý podle povahy stroje). Uvedení jazykového materiálu do stroje je prvním krokem SP a zároveň samostatným problémem. Nejčastěji je text převáděn do kódu stroje ručně (děrováním pásky apod.). Pro urychlení práce je však ne-

¹⁾ Sovětský lingvista I. I. Revzin se pro vzdálenou budoucnost domnívá, že bude možno k překladovým strojům, přidávat i rozsáhlé automatické encyklopedie, kde by podobné údaje mohl stroj zjistit sám.

zbytně třeba vypracovat automatické vstupy. U tištěného textu jsou práce již dosti daleko.²⁾ Zachycení mluvené řeči je neúměrně složitější (literary jsou standardní, ale výslovnost nikoli — zabarvení hlasu, individuální odchylky artikulace atd.).

Po uvedení textu do stroje začíná vlastní analýza; analýza textu je algoritmus,³⁾ jehož uplatnění na rozbíranou větu vede k tomu, že známe

- a) jednotky, z nichž je věta složena — analýza lexikální a frazeologická,
- g) gramatický tvar každé jednotky — analýza morfoloická,
- c) věcnou platnost každé jednotky — analýza syntaktická,
- d) přesný slovní význam každé jednotky — analýza sémantická.

Tyto čtyři složky nelze od sebe nijak přesně oddělit, analýza probíhá komplexně. Pro pochopení je však nutné pojednat o nich odděleně.

ANALÝZA LEXIKÁLNÍ A FRAZELOGICKÁ

Lexikální analýza může být přirovnána k vyhledání slov překládaného textu. Slovník stroje je ovšem podstatně jiný než slovníky běžné. Jde totiž o slovník kmenů (jedno slovo může mít ve slovníku více kmenů i údajů), který je uspořádan tak, aby vyhledávání bylo co možná nejefektivnější. Každý stroj nalezne obvykle pomocí odčítání⁴⁾ příslušné slovo, nalezne u něho též gramatickou informaci (slovní druh, vzor flexe, rod, u sloves vazbu apod.). Další údaj, zde uložený, je frazeologický index, který označuje, že slovo může být součástí sousloví, idiomu.

Již při překládání slovo za slovem bylo zjištěno, že sousloví, v němž jednotlivé složky ztrácejí svůj obvyklý význam, musí být překládány jako celek. Nelze např. po jednotlivých slovech překládat anglickou frázi *to take the floor* (vzít podlahu), poněvadž má jako celek význam „ujmouti

se slova“. Vedle slovníku jednotlivých slov je tedy ve stroji uložen též slovník frazeologický, v němž hledáme vždy, narazíme-li na slovo, které může být členem idiomu. Zda jím skutečně je, pozná stroj podle toho, že je s ním ve větě spojena další část idiomu.⁵⁾

Rozsah slovníku značně zatěžuje paměť stroje, a proto se alespoň pro začátek počítá s přípravou překladu omezenou na jistý obor, kde slovní zásoba je vždy značně zúžena. Nicméně i příprava takového oborového (elektronického, vojenského, lékařského apod.) slovníku je práce velmi náročná, poněvadž zvláště v oblasti terminologie a frazeologie chybí dosud systematický lingvistický rozbor. Také otázky činnosti slov a tvarů skýtají rozsáhlé možnosti výzkumu.

ANALÝZA MORFOLOGICKÁ

V morfoloické analýze vycházíme z gramatických údajů, získaných ze slovníku a zjišťujeme, v jakém tvaru to které slovo je, což je předpoklad pro určení jeho větné platnosti. K tomu je vypracován slovník koncovek, v němž identifikujeme část slova dosud nezpracovanou.⁶⁾ Vážným problémem je tu homonymie (tvarová víceznačnost), kterou odstraňujeme především porovnáním údajů ze slovníku kmenů a slovníku koncovek. Někdy se homonymie kmene a koncovky vzájemně ruší. Např.7)

Kmen: žen-

... sloveso	podstatné jméno
3. třídy	ženského rodu
	vzor 1

Koncovka: -ám

... sloveso	podstatné jméno
5. třídy	ženského rodu vz. 1
1. os. jedn. č.	3. pád mn. čísla

Tedy jde o 3. pád množného čísla podstatného jména žena. Někdy ovšem je homonymní tvar jako celek.

²⁾ Jednotlivá písmena se rozpoznávají např. počtem protěží různých čar jistého obrazce, který se překládá na obraz litery.

³⁾ Uspořádaný soubor pravidel, řešící jednoznačně daný úkol. Např. klíč k určování rostlin je algoritmus.

⁴⁾ Slovní kmeny jsou zaznamenávány v číselné podobě a srovnány tak, aby se čísla je zobrazující postupně zvyšovala. Stroj potom srovnává (např. odčítáním) postupně tato jednotlivá čísla se zadaným tak dlouho, až nalezne ve slovníku největší hodnotu obsaženou v zadaném slově. To je hledaný kmen, zbývající část zadaného slova je koncovka.

⁵⁾ Tyto otázky jsou nejvíce zpracovány pro angličtinu, kde je idiomů poměrně velmi mnoho.

⁶⁾ Dosud jsme pracovali toliko s kmenem, např. automobil —em, —y, —u atd.

⁷⁾ Neuvádíme zde celou bohatost homonymie uváděných prvků, jde jen o ilustraci.

Kmen: žen-

... sloveso	podstatné jméno
3. třídy	ženského rodu
	vzor 1

-u

... sloveso	podstatné jméno
3. třídy	ženského rodu
1. osoba	vzor 1
jedn. čísla	4. p. jedn. čísla

V takových případech musíme použít jiných ukazatelů, které jsou v textu formálně obsaženy, což je již otázka syntaktické analýzy. Hlavní problém analýzy a také těžiště přípravných prací je odstraňovat homonymii. Je třeba nově prozkoumat široký jazykový materiál, neboť v dosavadních gramatikách jsou podobné otázky rozebírány jen okrajově⁸⁾ a nesystematicky.

SYNTAKTICKÁ ANALÝZA

Syntaktická analýza je nejsložitější část algoritmu analýzy. Jejím úkolem je označit větnou platnost jednotlivých jednotek věty a určit jejich vzájemné vztahy. Stroj postupuje do určité míry obdobně jako žák při větném rozboru — vyhledává jednotlivé větné celky (tj. určí hranice mezi jednotlivými větami v souvětí) a v nich pak větné dvojice. Pomocí dotazů o okolí rozbíraného slova jsou dvojice vyhledány, což často řeší i různou zbyvajících homonymií. Např. stojí-li před tvarem „žena“ sloveso přechodné, které má předmět ve 4. pádě, jde o substantivum (*vidím ženu*), naopak následuje-li za tvarem *ženu* subst. ve 4. pádě, jde o přechodné sloveso (*ženu koně*). Někdy zůstávají celé fráze i věty víceznačné a řešení je třeba hledat v širším kontextu. V rovině syntaxe jsou všechny postupy daleko složitější. Zatímco ještě v morfologické analýze šlo převážně o vyhledávání v tabulkách a o porovnávání údajů, zde musíme použít k dosažení cíle často velmi složitý komplex operací.

SÉMANTICKÁ ANALÝZA

Ani úplná syntaktická analýza nedává ještě možnost překladu. Je nutné analyzovat věcný význam jednotlivých slov,

z nichž mnohá jsou opět homonymní. Zjišťujeme-li např., o který z významů anglického slova *track* (stopa, dráha, trať, sled) jde v právě rozbírané větě. Používáme prozatím dvou pomůcek. Prvou je omezení překladu na texty jednoho oboru, kde zpravidla podobné těžkosti z velké části odpadnou. Druhou je vyhledání typických spojení,⁹⁾ což je ovšem problém dosud málo zpracovaný. Případy, které nejsou zatím uspokojivě řešitelné, řešíme většinou tak, že jsou vytištěny obě eventualy, z nichž si čtenář vybere potřebnou podle svých zkušeností.

Výsledkem celkové analýzy je tedy věta zapsaná v řeči stroje. Její význam je zaznamenán jako soubor údajů sémantických, syntaktických, morfologických a lexicálních.

—☆—

Soubor těchto údajů je východiskem druhé etapy překladu, **syntézy** výstupního textu. Syntéza je algoritmus, který přiřazuje známému významu potřebné formy příslušného skutečného jazyka. Byla-li hlavním problémem analýzy homonymie, hlavním problémem syntézy je synonymie jazyka. Není totiž možné zcela jednoduše používat jen jednoho ze dvou (více) možných vyjádření. Souzračnost není nikdy stoprocentní a kromě toho existují (byť zatím mnohdy ne přesně popsání) pravidla pro výběr právě jednoho typu vyjádření. Popis těchto pravidel je právě lingvistický problém syntézy. Např. spojení *Miller's book* se do češtiny překládá *Millerova kniha*, ale spojení *professor Miller's book* se musí přeložit *kniha profesora Millera*.

**TEORETICKÁ ZÁKLADNA APLIKACÍ
— MODERNÍ JAZYKOVĚDA**

Z celkového popisu strojového překladu je jasné, že jeho uskutečnění si žádá rozsáhlý jazykový výzkum konkrétních jazyků, ale též mnohá otázky, které se objevily za 10 let trvání oboru především při spolupráci s matematikou a programátory. Celý soubor těchto otázek byl jedním z podnětů, které daly vznik nově orientovaným odvětvím jazykovědy, souborně označovaným jako **matematická lin-**

⁸⁾ Pro účely, kterým tyto gramatiky sloužily, jsou to totiž otázky nepodstatné.

⁹⁾ Např. *track of units* — sled jednotek, *track of plain* — dráha letadla apod. Uvádíme vědomě případ takový, kde řešení je pravděpodobné, ale ne zcela jisté — *track of plains* může být také sled letadel.

lingvistika. Je však třeba odlišovat tři její složky: lingvistiku algebraickou, kvantitativní a aplikovanou.

Lingvistika aplikovaná (přesněji aplikovaná matematická lingvistika) se zabývá především otázkami strojového překladu a též otázkami automatického referování, vyhledávání a ukládání informací, výhodného kódování pro telekomunikace a otázkami vyučování cizím jazykům (sem patří mj. otázky minimálních slovníků, které mají mnoho společného s přípravou SP). Lingvistika aplikovaná se rozvíjí na základě lingvistiky algebraické a kvantitativní, jímž naopak přináší mnohé závažné podněty.¹⁰⁾

Vztah teorie a praxe můžeme zde odlišovat. Proto také není úplně přesvědčivé tvrzení některých odborníků, že SP není uskutečnitelný a že je možné zabývat se jen teoretickými otázkami lingvistiky. Domníváme se, že budou-li tyto otázky vyřešeny, bude to předpoklad pro další rychlý vývoj SP. Je to stejně nutné jako dnešní stav, kdy mnoho odborníků, kteří narazili na neočekávané problémy při praktické přípravě SP, se stále více zabývá otázkami teoretickými, bez jejichž řešení nelze SP připravit. Domníváme se, že obě sféry práce je třeba spojovat a při teoretických problémech nezapomínat na možné a nutné aplikace výsledků, stejně jako nepostupovat ukvapeně (bez teoretického zájmu) v praktických přípravách SP, což by mohlo vést k vážným a zbytečným ztrátám.

Lingvistika kvantitativní se zabývá kvantitativní stránkou jazyka a přináší mnoho cenných poznatků jak pro lingvistiku samu, tak pro praxi. Viděli jsme, jak důležité jsou např. otázky frekvence slov pro SP. Velmi cenné jsou výsledky výzkumů opřených o matematickou statistiku, teorii pravděpodobnosti a nejnověji teorii informace.

Lingvistika algebraická vznikla v souvislosti se snahou popsat pomocí matematických metod systém jazyka. Úzký vztah mezi novou teorií gramatiky a matematickou teorií automatů vedl rychle k ustavení pomezí disciplíny, která se opírá na jedné straně o teorii rekurzivních funkcí, teorii automatů a matematickou logiku, na straně druhé o strukturní lingvistiku, zejména o dosavadní výsledky americké deskriptivní lingvistiky. V mnoha směrech znamená však algebraická

lingvistika odklon od deskriptivismu a návrat k některým hlediskům starším. Zdá se, že v tomto kontextu bude možno dobře využít určitých výsledků evropské strukturní lingvistiky a zejména pražské školy s jejím funkčním chápáním jazykových jednotek a jejich vztahů.¹¹⁾

Algebraická lingvistika usiluje o matematické zachycení jazykového systému, o formální vymezení jazyka. Jazyk je tu obvykle chápán jako množina gramaticky správných vět, kterou je možno různě vymežit. Důležité jsou dva způsoby: procedura **generativní**, která umožňuje vyčísřit (enumerate) všechny prvky množiny (pokud jde o množinu rekurzivně spočetnou), a procedura **rekognoskativní**, která umožňuje zjistit, zda určitý prvek je či není prvkem dané množiny (rekurzivní). Z lingvistického hlediska jsou na tomto poli nejzajímavější práce amerického badatele N. A. Chomského. Na generativní gramatiku klade dva požadavky:

— aby uplatňováním pravidel této gramatiky vznikaly právě jen všechny věty daného jazyka,

— aby posloupnost užití pravidel při vzniku věty (tzv. derivace věty) byla vhodným východiskem pro formální charakteristiku struktury této věty.

Vznikla celá řada gramatik, které byly porovnávány. Přitom se zřetelně projevila skutečnost, že v oboru pracují matematici a lingvisté. Matematiky zajímaly zejména otázky jednoduchosti gramatik a jejich generativní síly a byl objasněn jejich vztah k různým typům automatů.

Lingvisty zaujala otázka adekvátnosti gramatik pro popis daného jazyka. Za nejbližší přirozeným jazykům je obvykle považována gramatika transformační, která se z dosavadních typů nejvíce blíží vystižení našich intuitivních představ o jazyce, ale není zatím uspokojivě definována matematicky. Dosud ostatně nebyl připraven žádný úplný generativní popis některého jazyka,¹²⁾ což nepřekvapuje, neboť jde o úkol velmi rozsáhlý.

V poslední době se metody algebraické lingvistiky začínají uplatňovat i ve studiu sémantiky, jejíž matematické ovládnutí by jistě mělo dalekosáhlý význam pro aplikace jako je SP nebo automatické zpracování informací z textů.

¹¹⁾ Na vypracování tohoto přístupu se podíleli čeští lingvisté V. Mathesius, B. Havránek, V. Skalička a další.

¹²⁾ Nejrozsáhlejší jsou zatím práce R. B. Leese o angličtině a M. Bierwische o němčině.

¹⁰⁾ SP vyvolal nutně striktní postup od formy k funkci při popisu jazyka.

BUDOUCNOST SP, JEHO UPLATNĚNÍ A STAV PRACÍ U NÁS

Ukázali jsme, že SP staví před lingvisty a matematiky řadu velmi vážných problémů, které je nutné řešit. Jejich velikost a skutečnost, že s postupem prací dochází spíše k jejich komplikování než k hotovému řešení, vedla k tomu, že se znovu začalo uvažovat o možnostech SP vůbec. Názor našich odborníků, shodný s mnoha zahraničními odborníky, je takový, že SP je uskutečnitelný, ovšem etapově. Je nutné počítat jednak s překlady podle jednotlivých oborů, a to především těch, kde je vyjadřování značně standardizováno (technické obory vázané na matematiku), jednak s tím, že kvalita překladu se bude zvyšovat v souvislosti s řešením dnes ještě otevřených problémů a konečně i s rozvojem počítačů, což patrně přinese nové, větší možnosti.¹³⁾ Dnešní stav je takový, že přímé užití SP v dosažené kvalitě není ještě rentabilní. Známe jen některé dílčí způsoby užití (rovněž ovšem cenné). V Japonsku je v provozu stroj výborně překládající meteorologické zprávy, a to díky úzce vymezenému oboru. V USA překládá firma IBM dosti nedokonalým způsobem z ruštiny do angličtiny noviny s tím účelem, že zajímavé zprávy jsou pak zadány k lidskému překladu. V SSSR je připraven k provozu stroj pro překlad patentních informací. Poměrně v blízké době hodlají vstoupit do praxe pracovníci v Teddingtonu (Anglie). Zajímavý je také všeobecný zájem vojenských odborníků o otázky SP, což je dáno zřejmě možnostmi, které skýtají různé úzce vymezené obory vojenského odborného stylu. O přímém využití SP ve vojenství jsme ovšem beze zprávy.

Dnešní užití SP je možné jen u značně úzce vymezených oborů nebo za cenu malé kvality.

Pokusy plánované na různých místech však ukazují, že jsou připraveny překla-

dové systémy značně lepší než ty, které dosud známe. Uvědomíme-li si tyto skutečnosti a znovu zvážíme důvody, které vedou k myšlence automatizace překladu (nesmírné množství cizojazyčné literatury nezvládnutelné lidskými překladateli), musíme říci, že SP má velké možnosti a mohl by být velmi důležitý i tehdy, bude-li nutno text pro stroj upravit tzv. předredakcí (redaktor by určitým stylem text standardizoval) a výsledek korigovat postredakcí (redaktor by opravil hrubé nedostatky překladu).

Hlavní otázkou je rentabilita. I nedokonalý překlad strojem je dnes sám o sobě velmi drahý, je třeba počítat s odměnou případného postredaktora atd. Domníváme se však, že i tuto otázku musíme hodnotit z hlediska vývoje, který probíhá. Další výzkumy v oblasti samočinných počítačů povedou jistě ke zlevňování jejich provozu, v oblasti teoretické a aplikované matematické lingvistiky by mohly vést k přípravě tak dobrých překladů, že by práce postredaktorů mohla být postupně omezena na minimum.

V ČSSR je dnes vedle řady teoretických otázek (problematika převodního jazyka, teorie gramatiky ve vztahu k sémantice aj.) připravována analýza a syntéza češtiny (morfologické části byly již zkoušeny na počítači LGP-30 v Centru numerické matematiky MFF KU) a analýza angličtiny. Značně pokročily práce na přípravu anglicko-českého slovníku pro elektrotechnické texty a je připraven pokus anglicko-českého překladu (s využitím převodního jazyka), který bude ve spolupráci VÚMS a pracovišť Karlovy university uskutečněn na čs. počítači EPOS.

O úrovni oboru svědčí již bohatá literatura časopisecká (SaS, Slavica Pragensia), ohlas vystoupení čs. odborníků na mezinárodních konferencích, možnost založení vlastního samostatného Bulletinu, věnovaného otázkám matematické lingvistiky a jejich aplikací.¹⁴⁾

¹³⁾ V jednom z posledních příspěvků k diskusi o budoucnosti SP říká A. F. R. Brown, že otázka velikosti paměti, již dnes po zavedení fotoskopické paměti nehraje roli, a vypočítává, že i rychlost překladu (25 000 slov za hodinu) je dobrá.

¹⁴⁾ Prvé číslo bylo připraveno na Karlově universitě v září 1964.