Automatické zpracování vojenských jazykových informací

Major Miloslav Polanský, promovaný filolog

Na úseku automatického zpracování vojenských jazykových informací se řeší dva úkoly, a to statistický výzkum vojenského jazyka a strojový překlad.

Cílem prvního úkolu je postupně stanovit repertoár, rozložení četností a základní statistické charakteristiky prostředků vojenského jazyka (písmen a jejich digramových a trigramových kombinací, slov, syntaktických konstrukcí a vět) pro jednotlivé druhy vojsk, vojenské obory a velitelské stupně. Na základě statistického výzkumu vojenského jazyka, klasifikace operační, taktické a týlové informace a věcného posouzení vojenských operátorů a dalších vojenských odborníků lze uskutečnit kompresi a formalizaci velitelského jazyka a tím i formalizaci bojové a operační dokumentace. Statistický výzkum a formalizace vojenské jazykové informace jsou důležitým předpokladem pro efektivní kódování vojenských jazykových informací v samočinném počítači. Kompresí velitelského jazyka sledujeme zvýšit úspornost různých komunikačních systémů. Statistického výzkumu využije také kódovací a šifrovací služba ke zkvalitnění obsahové náplně slovníků šifer a hovorových tabulek tím, že do nich zařadí nejfrekventovanější ucelené fráze, rozkazy, hlášení, obraty, písmena a jejich kombinace, které vyjádří jedním kódovým znakem. Tak se zvýší jejich operativnost, kryptologická bezpečnost a hospodárnost.

V oblasti strojového překladu je snaha perspektivně pomocí řešit rozsáhlé úkoly vojenské a technické informační služby armády, tj. postupně v souladu s výzkumnými pracemi civilních institucí připravit po lexikografické a logickomatematické stránce cizojazyčné slovníky (klíče) pro vojenské a vojenskotechnické obory.

V otázkách statistického výzkumu vojenského jazyka spolupracujeme v rámci smíšené pracovní komise s oddělením matematické lingvistiky Ústavu pro jazyk český ČSAV.

Další články dvou význačných československých odborníků mají informovat čtenáře o otázkách matematické lingvistiky.

Koncepce statistického výzkumu vojenského jazyka

Doc. Dr. Lubomír Doležel, CSc, Ústav pro jazyk český ČSAV

V celé řadě úkolů, které dnes řešíme pomocí samočinných počítačů, zaujímá významné místo **automatické zpracování jazykových (nenumerických) informací**. Tím rozumíme takové operace, na jejichž vstupu je zadán nějaký jazykový text a na výstupu se objevuje určitá (v podstatě logická) transformace vstupního textu, např. překlad textu do jiného jazyka nebo kódu, vyhledání obsahových indexů, textu, résumé obsahu textu atd. Všechny tyto operace provádí zatím člověk. Moderní samočinné počítače umožňují předat tyto práce strojům a uskutečnit tak technickou revoluci v překládání, dokumentaci, referování, vyhledávání informací atd.

Automatické zpracování jazykových informací není možné bez exaktního studia

Mjr. Mil. Polanský, Doc. dr. L. Doležel, Doc. dr. P. Sgall, CSc. P. Piťha

kódu, v němž jsou tyto informace zobrazeny, tj. **přirozeného lidského jazyka.** Přirozený lidský jazyk je základním a univerzálním kódem nenumerických informací, i když přitom v některých oborech (zejména vědeckých a technických) může být ve větší nebo menší míře kombinován s různými umělými jazyky. Nutnost studovat kód jazykových informací spojuje problematiku automatického zpracování nenumerických informací s moderní jazykovědou.

Vývoj moderní jazykovědy určují do značné míry právě tyto její nové aspekty. Protože automatické zpracování jazykových informací vyžaduje naprosto exaktní rozbor jazykových faktů, pronikají do jazykovědy matematické a matematickologické metody a vzniká významný směr v moderní jazykovědě, tzv. **matematická lingvistika.**[1])

V současné matematické lingvistice vyhranily se dva dosti odlišné směry — tzv. statistická (kvantitativní) a algebraická (nekvantitativní) lingvistika. Zatímco v algebraické lingvistice se aplikují „nekvantitativní" obory matematiky (jako je např. topologie, moderní algebra apod.) k analýze a zobrazení **kvalitativně strukturních** vztahů mezi lingvistickými objekty, statistická lingvistika studuje **měřitelné,** kvantitativní vlastnosti jazykových objektů s použitím metod matematické statistiky, teorie pravděpodobnosti apod.

Oba uvedené směry mají v matematické lingvistice stejné oprávnění, protože každý má svůj specifický předmět výzkumu. V současném světovém vývoji můžeme pozorovat zřetelné tendence k syntéze obou přístupů. Rovněž problém automatického zpracování jazykových informací vyžaduje jak nekvantitativní, tak kvantitativní studium jazyka. To je prokázáno nejen teoreticky, ale také samým vývojem moderní jazykovědy i techniky zpracování jazykových informací.

Odborné předpoklady pro řešení otázek automatického zpracování jazykových informací platí rovněž pro specifickou oblast vojenských informací. Většina vojenských informací je uložena v jazykové formě a proto různé úkoly automatického zpracování vojenských informací nejsou řešitelné bez exaktního, kvalitativního i kvantitativního výzkumu vojenského jazyka. **Výzkum vojenského jazyka je tedy jeden ze základních předpokladů k uplatnění moderní automatizační techniky v armádě.** Musíme však varovat před krátkozrakým prakticismem, který vytyčuje úkoly lingvistickému zkoumání pouze z hlediska okamžitých dílčích technických úkolů. Úplná automatizace zpracování vojenských informací vyžaduje **soustavný, vysoce teoretický výzkum vojenského jazyka.** Takový výzkum se může zpočátku zdát málo efektivní z úzce praktického hlediska, avšak z hlediska řešení perspektivních úloh automatizace je nezbytný.

Cílem **statistického výzkumu** vojenského jazyka[2]) je zjistit repertoár jeho jazykových prostředků, jejich oblasti užívání, jejich četností apod. Tyto údaje jsou nezbytné pro jakékoli operace s vojenským jazykem a tedy také pro automatické zpracování vojenských informací. Statistický výzkum vojenského jazyka je založen na existujících zkušenostech statistické lingvistiky a zejména pak na zásadách, které jsou uplatňovány při statistickém výzkumu češtiny. Přitom ovšem není vyloučena eventuální modifikace postupů a metod, jakmile to bude nutné.

Základní otázky při statistickém výzkumu jazyka jsou:

1. Které vlastnosti jazykových objektů lze kvantitativně určovat (měřit)?
2. Na jakém materiálu se mají tyto vlastnosti zjišťovat?
3. Jaké statistické charakteristiky mají pro poznání jazykových objektů největší význam?

Pokusím se stručně odpovědět na jednotlivé otázky.

1. Pro statistickou lingvistiku je evidentní, že jazykové objekty vykazují kvantitativně měřitelné vlastnosti. Základními měřitelnými vlastnostmi jazykových jednotek (ponecháme-li stranou vlastnosti fyzikální, které charakterizují mluvený, řečový signál) jsou **četnost (frekvence)** a **rozsah (délka).** Četností jazykové jednotky rozumíme počet (absolutní nebo relativní) jejích výskytů v určitém textu, např. v četnosti určitého konkrétního slova. Důležité jsou rovněž údaje o četnosti tříd (stejnorodých skupin) jazykových jednotek, např. gramatických nebo lexikálních kategorií (četnost sloves nebo tvaru slovesného v určitém textu).

[1]) Poučný přehled dosavadních výsledků a přehled odborné literatury viz stať W. Platha Matematická lingvistika (český překlad ve sborníku Teorie informace a jazykověda, Praha 1964, str. 24—53.)

[2]) O některých otázkách (nestatistického) studia vojenské češtiny pojednává knížka K. Richtera a F. Švarce Vojenský jazyk a styl, Praha 1957.

Četnost jazykové jednoty zkoumáme obvykle spolu s četnostmi jiných jednotek téhož souboru, takže získáváme tzv. **rozložení četností**. Nejjednodušším příkladem takového rozložení jsou četnosti písmen abecedy. Pro češtinu jsme např. stanovili předběžně toto rozložení četností jednotlivých písmen:[3])

Grafém	Četnost	Grafém	Četnost
mezera	0,16586	Ň	0,01437
E	0,07261	B	0,01363
O	0,06866	H	0,01095
A	0,05431	É	0,01046
N	0,04036	C	0,01045
V	0,03953	CH	0,00974
T	0,03870	Ř	0,00971
S	0,03743	Ž	0,00956
K	0,03368	Ý	0,00852
I	0,03303	Č	0,00784
L	0,03293	Š	0,00746
U	0,02999	Ť	0,00652
R	0,02932	Ě	0,00619
P	0,02792	Ú	0,00546
M	0,02788	Ď	0,00414
D	0,02643	F	0,00194
Í	0,02490	G	0,00168
Á	0,02153	X	0,00062
J	0,02088	Ó	0,00042
Z	0,01902	W	0,00011
Y	0,01623	Q	0,00003
			1,00000

Z tabulky je na první pohled zřejmé, že jednotlivá písmena abecedy se vyskytují velmi rozdílně, rozložení četností je silně nerovnoměrné. Podobný obraz skýtá rovněž rozložení četností jiných souborů jazykových jednotek. Již tato skutečnost ukazuje, že by bylo velmi neefektivní přistupovat k technickým operacím s jazykovými jednotkami bez zřetele k zásadním rozdílům v jejich četnostech. Tyto rozdíly jsou dobře známy např. sdělovacím technikům; již W. Morse při sestavování své telegrafní abecedy respektoval alespoň intuitivně četnosti písmen anglické abecedy.

Rozsah (délka) jazykových jednotek není dosud ve statistické lingvistice plně doceněna, protože již existují pro některé jazyky výzkumy rozložení délky slov nebo délky vět. Délka jazykové jednotky není pouhým vnějším údajem, nýbrž může mnoho povědět o skladbě jednotky, o její výstavbě z jednotek nižších, jimiž se délka měří. Tak např. na základě údajů o délce slov je možno stanovit kvantitativní zákonitosti výstavby slov z nižších prvků, zejména slabik nebo morfémů. Délkou vět je pak možno vystihnout kvantitativní zákonitosti, jimiž se řídí obecné rysy syntaktické struktury textů. Rozložení délky vět je ve statistické stylistice jedním ze základních ukazatelů stylové osobitosti jazykového textu.

2. Lingvistické statistické údaje získáváme z jazykových textů. Porovnáním údajů pro různé texty můžeme stanovit třídy textů, které jsou si statisticky blízké nebo odlišné. Postupujeme tedy od analýzy textu k závěrům o celé skupině nebo skupinách textů, popřípadě o množině všech textů určitého jazyka. Takto pojatý statistický výzkum jazyka je induktivní.

Množina textů (i když určité specifické komunikační oblasti, jako je např. vojenský jazyk) je příliš rozsáhlá, než aby mohly být získány statistické údaje ze všech textů zkoumané množiny. Proto musíme nutně získat předběžný **výběr** textů, které budou sloužit jako vlastní materiál statistické analýzy. Výběr nemůže být libovolný, ale podle určitých kritérií, a to jednak jazykovědných, jednak statistických. Jazykovědná kritéria zaručují, že vybrané texty budou **typickými představiteli** zkoumané komunikační oblasti i její vnitřní diferenciace; statistická kritéria zaručují **statistickou reprezentativnost** výběru, tj. umožňují tvrdit, že statistické údaje, získané na vybraných textech, jsou dostatečně spolehlivým **odhadem** poměrů v celé zkoumané množině textů.

Nemáme-li o zkoumaném souboru textů vůbec žádných údajů, je třeba statistický výzkum uskutečnit ve dvou etapách:

Na omezeném výběru textů **předběžně odhadneme** potřebné statistické údaje, které nám slouží jako podklad k výpočtu rozsahu reprezentativního výběru se zadanou přesností. Pro výzkum vojenského jazyka jsme si v této etapě určili (podle lingvistických kritérií) dvacet typických textů, které představují různé oblasti vo-

[3]) Viz moji stať „Předběžný odhad entropie a redundace psané češtiny", Slovo a slovesnost 1963, č. 165—175.

Mjr. Mil. Polanský, Doc. dr. L. Doležel, Doc. dr. P. Sgall, CSc. P. Piťha

jenské komunikace. Podle statistických kritérií jsme pak získali **náhodný výběr** tímto postupem:

V každém vybraném textu jsme očíslovali (počínaje na libovolném místě) 1000 vět, z nichž jsme podle tabulky náhodných čísel vybrali 250 vět; náhodný výběr do jisté míry kompenzuje tlak tematiky, který statistické údaje zkresluje. Materiál pro předběžný odhad charakteristik vojenského jazyka představuje tedy 5000 vět.

Na reprezentativním výběru textů o přesně stanoveném rozsahu upřesníme potom **odhad hledaných statistických údajů.** Poněvadž půjde o materiál značně rozsáhlý, můžeme uskutečnit tuto etapu výzkumu pouze samočinným počítačem, adaptovaným k získávání lingvistických dat.

3. Statistická lingvistika usiluje především o získání takových údajů, které by vyjadřovaly specifické a přitom podstatné lingvistické vlastnosti jazykové výstavby textů. Protože takové vlastnosti textů nazýváme **vlastnosti stylové,** můžeme také nejrůznější statistické ukazatele, jež jsou kvantitativním vyjádřením těchto stylových vlastností, nazvat **stylové charakteristiky** (stylové ukazatele). Ve statistické lingvistice byla již definována řada stylových charakteristik; naším cílem bude získat hodnoty těchto charakteristik pro české vojenské texty. Tím se ovšem nevyhýbáme stanovení nových charakteristik, v statistické lingvistice dosud nedefinovaných; naopak, budeme zavádět nové charakteristiky, pokud to bude možné a potřebné.

V oblasti grafematické výstavby budeme zjišťovat ze známých charakteristik především **entropie soustavy grafémů.** Podle známého postupu C. Shannona, použitého již ve zmíněném odhadu entropie psané češtiny, získáme přibližně odhady rozložení četností pro jednotlivé grafémy i pro jejich kombinace (podle možností až do pentagrafů, tj. kombinací páté třídy) ve vojenských textech a vypočítáme příslušné hodnoty entropie pro každé získané rozložení. Srovnáním různých textových výběrů mezi sebou získáme podklad pro předběžné řešení otázky, zda se v rozložení četností písmen jednotlivé vojenské texty (anebo třídy textů) signifikantně liší nebo neliší.

Entropie grafémů, získaná z rozložení jejich četností, je základním údajem o jazykové stavbě textů a je nezbytným předpokladem pro jakékoli kódovací nebo šifrovací operace.

V oblasti analýzy slovníku zjišťujeme: **Rozložení délky slov** jednak v písmenech, jednak v morfémech nebo slabikách. Toto rozložení je výchozím údajem o skladbě slovníku vojenských textů a může také sloužit jako rozlišující znak jednotlivých textů. V souvislosti s tím bude možno získat repertoár morfémů (event. též slabik) s hodnotami jejich četností. Také tyto údaje mají značný význam pro kódování operace, zejména pro některé metody automatického kódování nebo automatické komprese.

Frekvenční slovník jednotlivých slov zkoumaných textů, tj. seznam všech **slovních tvarů** daného textu s příslušným údajem četnosti. Je základním materiálem lexikální statistiky a můžeme z něho čerpat jednak další údaje o statistických poměrech ve slovní zásobě zkoumaných textů, jednak je nutným předpokladem jakýchkoli operací, směřujících k optimalizaci slovníku.

Poměr type/token, tj. poměr počtu různých slov a k počtu všech slov. Tato charakteristika nám prozrazuje velmi mnoho o bohatství anebo stereotypnosti slovníku, protože v podstatě udává průměrný počet opakování jednoho slova v textu. Hodnota této charakteristiky je ovšem závislá na délce (rozsahu textu) a proto je třeba buď fixovat délku textu, nebo provést určité numerické transformace v poměru type/token.

Busemannův koeficient jako poměr sloveso/adjektivum. Tato charakteristika slouží k rozlišení textů svou podstatou dějových, akčních a textů popisných, kvalitativních (vyjadřují především vlastnosti předmětů). Vedle Busemannova koeficientu je možno zkoumat rozložení všech slovních druhů, zejména poměr sloveso/jméno.

V oblasti syntaktické stavby zjišťujeme **rozložení délky vět.** Toto rozložení je jeden z nejvýraznějších ukazatelů stylu, protože užívání dlouhých nebo krátkých vět je jedním ze základních činitelů jazykové výstavby textů. Podstatné rozdíly v uvedeném rozložení je možno doložit již z hodnot průměrné délky věty; pro pět různých textů psané češtiny jsme např. zjistili tyto průměry (vyjádřeno v počtu slov): $\overline{X_1} = 4{,}86$, $\overline{X_2} = 9{,}54$, $\overline{X_3} = 10{,}20$, $\overline{X_4} = 14{,}12$, $\overline{X_5} = 18{,}06$. U výběru z vojenského textu (Vševojsk-1-2) jsme zjistili průměrnou délku věty 27,13. Již tento

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 nejzákladnějších 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. Upozorňuji zejména na úkol optimalizace, resp. **minimalizace** vojenského jazyka. Minimalizací vojenského jazyka rozumíme snížení repertoáru jazykových prostředků na nejnižší 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 řídké 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. Piťha,
Matematicko-fyzikální fakulta KU

Strojový překlad můžeme do určité míry chápat jako model činnosti překladatele. V prvém 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 zapitřebí.[1]) Jinou odlišností je nedostatek lidské paměti, která lidskému překladateli umožňuje řešit opakující se úlohy přímo, zatímco stroj musí vždy projít všemi body zdlouhavé analýzy znovu. Tento rozdíl však nelze chápat jako nevýhodu strojů, uvážíme-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 prvým 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.