1.3 Počítačová lingvistika

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (Czech). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 12–14.

Persistent URL: http://dml.cz/dmlcz/402315

Stále ale není zcela jasné, zda vůbec lze v úplnosti formálně zachytit tak složitý systém, jakým bezesporu přirozený jazyk je. Zatím se zdá, že gramatiku jazyka (fonologie, morfologie, syntax) je možno formálními prostředky celkem úspěšně popsat, i když se jedná o oblast poměrně složitou. Dosud zatím vyčerpávajícímu formálnímu popisu odolává sémantika jazyka. Tento úkol je bezesporu náročný, ale vyžaduje jej sama praxe zejména v souvislosti s rozvojem výpočetní techniky a tvorbou různých počítačových programů. Ale to už se dostáváme k dalšímu odvětví matematické lingvistiky, a to k lingvistice počítačové.

1.3 Počítačová lingvistika

Je třetím tradičně rozlišovaným odvětvím matematické lingvistiky. Název *počítačová lingvistika* nám říká, že se jedná o počítačové zpracování jazyka, jež bylo prováděno zpočátku na jednoduchých děrnoštítkových strojích (podle nich také starší označení *strojová lingvistika*), později na složitých počítačích (odtud *počítačová* nebo také někdy *komputační lingvistika*).

Počítačová lingvistika se vyvíjí od konce padesátých let minulého století, a to zejména v souvislosti s rozvojem kybernetiky, výpočetní techniky, kvantitativní a algebraické lingvistiky a jiných hraničních oborů.

Připomeňme si postavení počítačové lingvistiky v rámci matematické lingvistiky: matematická lingvistika je tvořena dvěma teoretickými obory, a to lingvistikou kvantitativní a lingvistikou algebraickou. Počítačová lingvistika je potom jejich praktickou aplikací, proto se můžeme setkat i s termínem *aplikovaná matematická lingvistika*. Důležité je v tomto případě slovo „matematická“, neboť názvu aplikovaná lingvistika by odpovídala oblast podstatně širší, a to nejrůznější aplikace jazykovědy – jazykové vyučování, kultura spisovného jazyka, uplatnění jazykovědných výsledků v jiných disciplínách (např. literární věda, historie) atd.

Mezi hlavní problémy řešené v rámci počítačové lingvistiky patří strojový překlad, automatický rozbor, uchovávání a vyhledávání informací, vytváření jazyků pro automatické programování, v současnosti pak zejména tvorba nástrojů pro počítačové zpracování přirozeného jazyka. Jako součást počítačové lingvistiky je chápána i *korpusová lingvistika*, která pracuje s *korpusy*, tj. rozsáhlými soubory jazykových dat.

Mohutně rozvíjet se začala počítačová lingvistika především v souvislosti se *strojovým překladem* (rozumíme jím převedení textu ze vstupního jazyka do jazyka výstupního, cílového, pomocí stroje). Od 50. let 20. století na přípravě strojových překladů pro různé jazyky pracovaly desítky pracovišť po celém světě (první pokusy byly provedeny v roce 1954 v USA a roku 1955 v SSSR). Nakonec se ukázalo, že se jedná o úkol značně složitý. Pro účel překladu nestačí totiž jen zvládnutí mluvnické stavby, ale je třeba zakomponovat i sémantiku přirozených jazyků. Dosud nebyla sémantika zpracována natolik dostatečně, aby mohla být zachycena formálními metodami. Navíc rozsáhlá slovní zásoba a složitá stavba jazyků kladou značné nároky na paměť počítače i na dobu zpracování. Poměrně zdařilé se jeví automatické překladače určité speciální skupiny

textů (technické texty apod.), kde je poměrně úzká a stabilní terminologie a kde nezáleží na stylistických odstínech. Nejlepší výsledky zatím vykazují překladače poloautomatické, ve kterých jsou sporné jevy dodatečně redigovány člověkem.

Vyhledávání informací je úkol poměrně jednoduchý a poradily si s ním už děrnoštítkové stroje. *Automatické informování o obsahu* je záležitost mnohem náročnější. Kvalitní referát by měl být totiž jasný a srozumitelný, měl by být stručným obsahem práce, uvádět hlavní výsledky práce, její přínos a hodnocení. V zásadě se využívají dva postupy. První z nich využívá statistiky. Pro daný obor se určí soubor nejdůležitějších termínů a do referátu se automaticky přejímají ty věty, které těchto termínů obsahují nejvíce. Ze zkušenosti ale víme, že termíny nejčastěji se vyskytující nemusí být vždy nejvýznamnější a že důležité pasáže nemusí takový termín obsahovat vůbec. Druhý postup je založen na hledisku sémantickém, ale zde narážíme na ještě větší problémy, neboť bez důsledné formalizace stránky sémantické lze pracovat jen velmi obtížně. Nejnáročnějším stupněm z postupů zpracovávajících informace pomocí počítače jsou *informační jazyky*, tzn. ucelené teoretické systémy sloužící k ukládání a vyhledávání informací. Snaží se řešit řadu značně složitých úkolů – například je žádoucí, aby byl počítač schopen zodpovědět určitou otázku týkající se problematiky daného oboru na základě souboru poznatků z literatury tohoto oboru; aby odpovídal na otázky kladené v přirozených jazycích; doplňoval informace, které nejsou v textu explicitně vyjádřeny; zpracovával literaturu v různých jazycích apod.

Velkým pomocníkem nejen při tvorbě strojových překladů by se mohla stát *korpusová lingvistika*, která zaznamenává v současnosti obrovský rozvoj. Korpusovou lingvistikou rozumíme tu část počítačové lingvistiky, která se zabývá tvorbou a využitím jazykových *korpusů*, tzn. souborů jazykových dat. Tento soubor může mít podobu textů, které jsou zachyceny na papíře (nejčastěji v podobě excerpt), nebo může mít podobu elektronickou, kdy jsou jazyková data uložena v počítači. Korpus lze potom chápat jako rozsáhlý, vnitřně uspořádaný a ucelený soubor jazykových dat, která jsou elektronicky uložena, zpracována a přístupna. Ačkoliv se různých jazykových korpusů uchovaných na papíře užívalo velmi dávno (vzpomeňme na J. A. Komenského, kterému při požáru v Lešně shořel celý jeho rozsáhlý materiál na připravovaný latinsko-český a česko-latinský slovník), o skutečné korpusové lingvistice mluvíme až v souvislosti s rozvojem počítačové techniky.

Vztah počítačové lingvistiky a lingvistiky klasické je oboustranný. Na jedné straně pomáhá jazykověda ostatním oborům (zejména matematice a logice) řešit úlohy, které s lingvistikou přímo nesouvisí (např. programovací jazyky). Na straně druhé lze počítače využívat přímo v lingvistické práci (z těch jednodušších například třídění a statistické zpracování rozsáhlých souborů jazykových dat, z těch složitějších strojový překlad, automatické generování vět podle určitých pravidel, automatický rozbor větných struktur, automatické vyhledávání textových informací – rešerše aj.). Oblast počítačové lingvistiky je ale širší, zahrnuje totiž i vlastní jazykovědný výzkum, zejména z oblasti algebraické lingvistiky (počítače se využívají k prověřování různých gramatik). Spolu s lingvistikou algebraickou a kvantitativní klade počítačová lingvistika

otázky, které si tradiční lingvistika neklade, a napomáhá tak odhalovat nové stránky přirozených jazyků.