1.2 Algebraická lingvistika

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (Czech). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 9–12.

Persistent URL: http://dml.cz/dmlcz/402314

v polovině 19. století, který se pokoušel o nalezení pravidelností a zákonitostí v hláskové stavbě slovanských slov. V kapitole 2.11 si blíže představíme dvě statistiky, které vyšly již roku 1831 v časopise *Krok*. V roce 1886 vychází článek [57] matematika, fyzika a astronoma Augustina Seydlera, ve kterém se pomocí počtu pravděpodobnosti snaží dokázat nepravost *Rukopisů*[7]. Za předchůdce kvantitativní lingvistiky na našem území můžeme rovněž považovat některé členy Pražského lingvistického kroužku, zejména Viléma Mathesia, který se zabýval *potencionálností* jazykových jevů. V jazyce neplatí „absolutní zákony", ale v řeči každého jednotlivce existuje jisté kolísání „v určitých mezích a s určitou tendencí". Tyto tendence jsou pak podle Mathesia statisticky postižitelné. Kvantitativní lingvistika jako samostatná disciplína se u nás rozvíjela v rámci strukturalismu a je spojena s *pražskou školou* (J. Vachek, J. Krámský, B. Trnka, M. Těšitelová aj.).

Z tohoto stručného historického přehledu je zřejmé, že kvantitativní lingvistika měla praktický význam pro celou řadu oblastí. Zpočátku se jednalo o takové oblasti jako těsnopis, výuka pravopisu, šifrování, výuka cizích jazyků apod. Vladimír Šmilauer v [63] například uvádí zajímavou myšlenku, podle níž by se pro potřeby vyučování češtiny pro cizince sestavil katalog doporučených knih, které by byly srovnány podle míry *koncentrace slovníku*[8]. Až později kvantitativní lingvistika napomohla k lepšímu poznání různých jazykových jevů. Může být rovněž velkým pomocníkem v otázkách sporného autorství, překladatelství, srovnávací lingvistiky apod. A zejména v posledních letech se výsledky kvantitativní lingvistiky využívají při sestavování různých programů pracujících s přirozeným jazykem v rámci počítačové lingvistiky.

1.2 Algebraická lingvistika

Algebraickou lingvistikou rozumíme tu část matematické lingvistiky, která využívá nekvantitativních matematických metod, jako jsou algebra, teorie grafů, matematická logika, topologie, teorie množin či kombinatorika.

Algebraická lingvistika se začala formovat v druhé polovině 50. let 20. století zejména v souvislosti s potřebami *strojového překladu*. Ukázalo se totiž, že zatím jediným možným způsobem, jak odstranit nedostatky strojového i teoretického jazyka, je jeho důsledná formalizace. Autorství termínu *algebraická lingvistika* je připisováno Y. Bar-Hillelovi. U některých matematiků a lingvistů bývalého Sovětského svazu se můžeme setkat také s označením *teorie jazykových modelů*.

Společně s lingvistikou kvantitativní tvoří algebraická lingvistika teoretické obory matematické lingvistiky. Lingvistika počítačová je pak jejich praktickou aplikací.

Základ algebraické lingvistiky tvoří zejména tyto teorie: *generativní a transformační gramatika* N. Chomského, *rekognoskativní a kategoriální gramatika* Y. Bar-Hillela, *aplikačně generativní model jazyka* S. K. Šaumjana, *analytické*

[7] Podrobněji viz kap. 2.12.

[8] Veličina vyjadřující poměr prvních 50 nejčastějších (eventuálně 10 nejčastějších) plnovýznamových slov k délce textu. Viz též kap. 2.12.5.

modely jazyka (O. S. Kulaginová, R. L. Dobrušin, A. N. Kolmogorov, S. J. Fitialov, A. V. Gladkij, I. I. Revzin, C. Crăciun, S. Marcus, L. Nebeský aj.), *závislostní gramatika*, u nás např. *funkční generativní popis jazyka* (P. Sgall a kol.).

Významným mezníkem pro rozvoj algebraické lingvistiky bylo vydání knihy amerického lingvisty Noama Chomského *Syntactic structures* (1957)[9], ve které navrhl první variantu *mluvnice* nazývané *generativní* nebo *transformační* (možné jsou i oba názvy). Generativní znamená to, že jazyk je chápán jako tvůrčí proces, ve kterém se jednotlivé věty generují podle určitých pravidel, kterých existuje (stejně jako jednotek jazyka) omezený počet. Umožňují ale generovat neomezené množství vět. Při pokusu formalizovat popis gramatiky se snažil o jeho maximální zjednodušení. Zavádí proto termín *jádrových vět* (*kernel sentences*), to znamená základních jednoduchých vět, z nichž jsou všechny ostatní věty a souvětí odvozeny pomocí *transformačních pravidel*. Odtud tedy označení *generativní transformační mluvnice*. Protože první varianta mluvnice byla čistě formální (neuvažovala jazykový obsah), přišel N. Chomsky po četných námitkách v roce 1965 a letech následujících s variantou druhou, která měla již zakomponovanou složku sémantickou. Generativní transformační gramatika N. Chomského měla celou řadu následovníků, z nichž si pozornost zaslouží zejména Jerry Alan Fodor a Jerrold Katz, kteří se roku 1963 rovněž zasloužili o zahrnutí složky sémantické. O zakomponování pragmatické složky se poprvé v roce 1971 zmiňuje Yehoshua Bar-Hillel. U nás byl průkopníkem generativní gramatiky především Petr Sgall spolu se svými spolupracovníky, kteří vypracovali tzv. *funkční generativní popis jazyka*[10].

Opakem gramatiky generativní je *gramatika rekognoskativní*. Tento typ gramatiky vychází z konkrétní věty jazyka. Větu převádí na řetěz symbolů, odhaluje strukturu této věty a dále to, zda se jedná o gramaticky správnou větu příslušného jazyka. Nejznámějším příkladem rekognoskativní gramatiky je teorie vypracovaná na základě dřívějších systémů Yehoshuou Bar-Hillelem. Podle klíčového pojmu *kategorie* bývá tato gramatika nazývaná *kategoriální gramatikou*. Protože je kategoriálních gramatik celá řada (sám Bar-Hillel navázal zejména na práce K. Ajdukiewicze a H. B. Curryho), bývá Bar-Hillelova gramatika označována někdy termínem *kategoriální mluvnice identifikačního typu*[11].

V Sovětském svazu vytvořil na začátku 60. let 20. století S. K. Šaumjan gramatickou teorii s názvem *aplikačně-generativní model*, který v sobě zahrnoval prvky strukturalismu a generativní mluvnice. V Šaumjanově gramatické teorii jsou jazykové jednotky označeny symboly a odvozují se pomocí tzv. *aplikace* (metody matematické logiky, která se týká vztahů mezi symboly). Jazykověda by neměla zkoumat pouze jevy bezprostředně pozorovatelné, ale zejména hlubší souvislosti, které jsou za nimi skryté (roviny tzv. *logických konstruktů*). Důraz by měl být kladen na takové jevy, které jsou společné všem jazykům. Systém těchto univerzálních jevů nazývá Šaumjan *genotypický jazyk*. Naproti tomu *jazykem fenotypickým* rozumí ty jevy, které se vyskytují v jednotlivých

[9] Česky ji vydalo nakladatelství Academia v Praze roku 1966.
[10] Podrobně viz [58], [6].
[11] Více viz [58], [6].

přirozených jazycích. Jeho model se tedy skládá ze dvou částí, genotypické a fenotypické. První z nich je obecná teorie jazykových univerzálií a druhá je generativní mluvnice jednotlivých přirozených jazyků.

Generativní a rekognoskativní gramatiky byly zkoumány zejména na angličtině a ukázalo se, že dobře vyhovují pro popis jazyků s pevným slovosledem a jednoduchou morfologií, což angličtina právě je. Pro jazyky flexivní (zejména jazyky slovanské) jsou však tyto gramatiky nevyhovující. Zejména sovětští gramatici na tyto nedostatky upozorňovali a snažili se vytvořit obecnější formu gramatiky vyhovující typologicky různým jazykům. Tak vytvořili tzv. *analytické modely jazyka* jako protiklad k *modelům syntetickým* (rekognoskativní a generativní mluvnice). Zatímco v syntetických modelech vytváříme soubor gramaticky správných vět jazyka či zjišťujeme, které věty do tohoto souboru patří, v analytických modelech postupujeme opačně. Výchozím pojmem je pro nás soubor gramaticky správných vět jazyka. Za zakladatelku tohoto směru je považována O. S. Kulaginová, k rozvoji této teorie přispěli v 60. letech 20. století zejména matematici R. L. Dobrušin, A. N. Kolmogorov a lingvisté S. J. Fitialov, A. V. Gladkij a I. I. Revzin. Teorie analytických modelů je založena na teorii množin a jedná se o první využití teorie množin v lingvistice. Z bývalého východního bloku se o rozvoj tohoto modelu zasloužili rumunští vědci Solomon Marcus, jeden z nejvýznamnějších teoretiků analytické metody, a C. Crăciun. U nás se analytickou metodou zabýval především L. Nebeský.

Mezi další vědce, kteří se zasloužili o rozvoj algebraické lingvistiky v Sovětském svazu, patří například O. S. Achmanova, N. D. Andrejev, R. M. Frumkina, I. A. Melčuk, J. V. Padučeva aj.

Teorie množin byla využita vedle teorie analytických modelů rovněž v celé řadě různých lingvistických oblastí. V článku [49] můžeme najít několik takových zajímavých aplikací. Například Harary a Paper (1957) využili několika pojmů z teorie relací pro popis distribuce fonémů. Tento systém vedle strukturní charakteristiky vykazuje rovněž charakteristiky kvantitativní. Autoři zavádí indexy, které navrhují aplikovat na typologický výzkum distribuce fonémů. Lingvisté *glosematické školy* (např. Louis Hjelmslev) se pokusili vytvořit tzv. *kalkul nekvantitativních funkcí*, novou algebru zaměřenou zejména na popis „humanitních“ materiálů, např. jazyka. Tato *glosematická algebra* používá převážně terminologie logické teorie tříd, ale s tím důležitým rozdílem, že některé známé pojmy, jako např. funkce a negace, jsou definovány neobvyklým způsobem. Kritika Ungeheuerem ukázala, že Hjelmslev a Uldall nepoužívají glosematické terminologie vždy důsledně a vývoji jejich algebry by prospěl propracovaný logický rámec.

Algebraická teorie gramatiky je oblast, kde se matematika a lingvistika spojují nejtěsněji. Teorie gramatiky má totiž celou řadu společných znaků s matematickou *teorií automatů*. Z tohoto hlediska lze gramatiku jazyka chápat jako soustavu pravidel, které vymezují všechny správně tvořené věty daného jazyka. Pravidla takové gramatiky je možno zachytit formálně podobně jako pravidla vymezující činnost automatu. Různé typy automatů a gramatik lze srovnávat. Ukazuje se ovšem, že jednoduché typy gramatik nejsou pro lingvistiku dostatečné. Podle dosavadních výsledků nejvíce vyhovují gramatiky transformační.

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