Matematická lingvistika

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (Czech). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 5–7.

Persistent URL: http://dml.cz/dmlcz/402312

Kapitola 1

Matematická lingvistika

Termín *matematická lingvistika* můžeme chápat dvěma způsoby. Jednak jím lze označovat hraniční disciplínu mající stejně blízko k matematice i jazykovědě, jednak jím rozumíme tu část lingvistiky, která využívá matematických metod. Zpočátku převládalo pojetí druhé, tj. matematická lingvistika byla považována za lingvistickou disciplínu využívající matematických metod. Později se však matematická lingvistika začala chápat jako pomezní disciplína stojící mezi matematikou a lingvistikou, nověji i *umělou inteligencí* (AI – *artificial inteligence*) a *informatikou*. Cílem takto pojímané matematické lingvistiky je exaktní popis přirozeného jazyka opřený o matematické metody.

V současnosti se matematická lingvistika standardně rozděluje na tři odvětví podle toho, kterých matematických metod využívá, a to na:

1) ***lingvistiku kvantitativní*** (podle převládající metody bývá též označovaná jako *lingvistika statistická*), která pro studium přirozeného jazyka využívá kvantitativní matematické metody, jako je například matematická statistika, teorie pravděpodobnosti aj.;

2) ***lingvistiku algebraickou*** (v současnosti je rozšířenější označení *formální lingvistika*) využívající nekvantitativní matematické metody (logika, algebra, teorie množin, teorie grafů aj.) a zabývající se formálním popisem gramatik a jazyků;

3) ***lingvistiku počítačovou*** (též *komputační*, starší název *lingvistika strojová*), která ke studiu jazyků využívá počítače (dříve např. děrnoštítkové stroje) a metody informatiky. Dnes se můžeme setkat také s termínem *počítačové zpracování přirozeného jazyka* (*natural language processing* – NLP). Blízko k počítačové lingvistice má samostatně stojící *language engineering*, pod kterým si můžeme představit algoritmické techniky pro popis přirozeného jazyka a různé softwarové nástroje vzniklé jejich implementací, jež pak hrají roli v problematice např. strojového překladu, automatického rozboru, uchovávání a vyhledávání informací, tvorbě gramatických korektorů apod. Od začátku 90. let 20. století se rovněž prudce

rozvíjí tzv. *korpusová lingvistika*, tj. ta část lingvistiky, která zkoumá jazyk za pomoci rozsáhlých souborů jazykových dat (korpusů) uchovávaných zpravidla elektronicky.

O matematické lingvistice jako o samostatné disciplíně se mluví zhruba od konce 50. let a začátku 60. let minulého století. Zpravidla se jako počátek označuje rok 1957, kdy se konal VIII. mezinárodní lingvistický kongres v Oslo. Jak ale zmiňuje W. Plath v [49], můžeme aktivitu v této oblasti vysledovat již dříve, zejména v Americe a v Evropě, ale i na Dálném východě[1]. Profesor Joshua Whatmough se ve své zprávě na konferenci v Oslu zmínil o semináři matematické lingvistiky na Harvardově univerzitě v roce 1955, čímž se tento obor vůbec poprvé objevil v učebním programu vysoké školy. V dalších letech se tyto kurzy matematické lingvistiky rozšířily vedle Harvardovy univerzity i na jiné univerzity v USA (například na Massachusettskou vysokou školu technickou, univerzity v Michiganu a Pennsylvanii), v Německu na univerzitu v Bonnu a v Sovětském svazu byly kurzy matematické lingvistiky vyučovány na moskevské a leningradské univerzitě. Kurzy kvantitativní lingvistiky byly zavedeny na univerzitách v Indianě a Kalifornii v USA. V roce 1958 vydal sovětský ministr školství nařízení, aby pro studenty matematiky a filologie zavedli rektoři univerzit v Moskvě, Leningradě, Gorkém, Saratově a Tomsku volitelné přednášky o strojovém překladu a matematické lingvistice.

Od počátku se projevovala v rámci matematické lingvistiky jistá terminologická nejednotnost (např. vedle označení kvantitativní lingvistika se setkáváme rovněž s označením lingvistika statistická). Tato nejednotnost odrážela vlastně hledání ve vymezení matematické lingvistiky jako celku, ale i jejích složek. Znovu se tato problematika oživila počátkem 60. let 20. století, kdy se utvářel pojem matematické lingvistiky i její terminologie[2]. Na IX. mezinárodním kongresu lingvistů v americké Cambridgi v roce 1962 vznesl námitky proti termínu matematická lingvistika (resp. i proti pojetí matematické lingvistiky, jak se v 60. letech 20. století vyvíjelo) dánský matematický lingvista H. Spang-Hansen[3]. Navrhl rozlišovat lingvistiku kvantitativní (*arithmetical and statistical*) a nekvantitativní (*non-quantitative*), které mohou být obě jak nestrukturální, tak strukturální (kde strukturální lingvistiku chápe jako axiomatický popis kvalitativní stránky jazykových jevů).

Matematická lingvistika se nesnaží nahradit lingvistiku „klasickou“. To ani není vzhledem k povaze jazyka dost možné, neboť v jazyce neexistují pouze rysy kvantitativní, ale i rysy kvalitativní. Vladimír Šmilauer v [63] uvádí názorný příklad přibližující rozdíl mezi lingvistikou kvantitativní a kvalitativní. „*Protože*“ a „*poněvadž*“ jsou z hlediska lingvistiky kvalitativní spojky pod-

[1] V Japonsku vznikla v prosinci roku 1956 jako první na světě společnost *The Mathematical Linguistic Society of Japan* a v roce 1957 byl založen časopis *Mathematical Linguistics*. Více viz Köhler, Reinhard – Altmann, Gabriel – Piotrowski, Raimund G.: *Quantitative linguistics*. Walter de Gruyter, 2005.

[2] Viz Novák, P.: *K vytváření terminologie matematické lingvistiky*. Čs. terminologický časopis, č. 4, 1963, s. 234–237.

[3] Viz Spang-Hansen, H.: *Mathematical linguistics – a trend in name or in fact?* Preprints of Papers for the Ninth International Congress of Linguistics, Cambridge, Mass. 1962, s. 133–138.

řadicí příčinné. Z hlediska lingvistiky kvantitativní víme, že spojka „*protože*“ se z celkového počtu 1 623 527 slovních výskytů doložených ve FSČ[4] vyskytuje v počtu 1224, a to v 63 textech (ze 75 možných) a ve všech fukčních stylech. Nejvíce dokladů má v beletrii (477 výskytů), v literatuře pro mládež (263 výskytů) a v dramatech (152 výskytů), v pořadí slov podle jejich častosti je na 134. místě. Naproti tomu spojka „*poněvadž*“ má jen 374 dokladů v 19 dílech (textech), a to zejména v literatuře odborné (153 výskytů) a vědecké (107 výskytů), v básních není doložena vůbec (na rozdíl od spojky „*protože*“, která se v poezii vyskytuje 489krát). Lze tedy na závěr říci, že „*protože*“ je spojka běžná, „*poněvadž*“ je méně obvyklá a vyskytuje se především v próze naukové.

Cílem matematické lingvistiky je exaktní popis přirozeného jazyka opřený o matematické metody. Matematická lingvistika se snaží hledat nové otázky či problémy, dále exaktními metodami potvrdit výsledky, kterých dosáhla jazykověda bez užití matematických metod. Může být rovněž pomocníkem jiných oborů – např. sdělovací technika, automatizace či samotná matematika. Je potřeba se ale vyhnout nekritickému zavádění matematického aparátu, které by vedlo k nic neříkajícím výsledkům. Zjištěná data je nutno vždy převést do řeči té disciplíny, na kterou je matematický aparát aplikován, zde tedy do řeči lingvistiky. Proto by měli spolupracovat navzájem lingvisté i matematici, popř. informatici.

Často se objevují námitky, zda je vůbec možné matematickými prostředky popsat tak složité systémy, jakými přirozené jazyky bezpochyby jsou. Pravdou ale je, „*že jakmile se podaří nějaký úsek skutečnosti přesně popsat a jeho obecné zákonitosti vědecky zachytit, pak může být zpracován i matematicky. Hranice těchto možností jsou tedy dány především dosavadním stavem té které vědy (. . .)*“ ([58], s. 73). Největší problémy, se kterými se setkáváme při popisu přirozeného jazyka, jsou spojeny zejména s významovou vágností ve všech jazykových rovinách a dále s velkou mírou synonymie a homonymie. Proto je jedním z hlavních cílů matematické lingvistiky v současnosti tvorba nástrojů pro automatické odstraňování víceznačných interpretací jednotek jazyka (*automatická desambiguace*). Jeho vyřešení má pak dalekosáhlý význam pro rozvoj veškerých aplikací v oblasti matematické lingvistiky.

1.1 Kvantitativní lingvistika

Tímto termínem označujeme tu část matematické lingvistiky, které využívá kvantitativních matematických metod. Jsou to například statistika, matematická statistika či teorie pravděpodobnosti. Mocným impulsem se pro rozvoj kvantitativní lingvistiky staly práce C. E. Shannona a N. Wienera z konce 40. let 20. století, které položily základy matematické *teorie informace*. Tato teorie se zabývá kvantitativními vlastnostmi sdělovacích soustav, má však také závažné praktické aspekty týkající se sdělovací techniky (úspornost kódování, odolnost kódů proti chybám, šumu apod.). Národní jazyky jako nejdůležitější

[4] Běžně užívaná zkratka pro [25].