1.1 Kvantitativní lingvistika

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (Czech). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 7–9.

Persistent URL: http://dml.cz/dmlcz/402313

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].

sdělovací soustavy lidské společnosti se tedy zcela přirozeně staly centrem pozornosti matematiků.

Podle převažující metody bývá v literatuře někdy lingvistika kvantitativní nazývána *statistická lingvistika*[5]. Jak již bylo zmíněno dříve, tato terminologická nejednotnost odráží hledání předmětu matematické lingvistiky jednak jako celku, jednak jejích jednotlivých složek, tedy i lingvistiky kvantitativní. Protože je statistika nejčastější matematickou metodou uplatňovanou v rámci kvantitativní lingvistiky, můžeme se setkat rovněž s termíny jako *fonologická statistika*, *morfologická statistika*, *lexikální statistika*, *stylistická statistika* apod., které vypovídají o tom, která část lingvistiky tyto statistické metody využívá.

Kvantitativní lingvistika má ze všech tří odvětví matematické lingvistiky (vedle lingvistiky algebraické a počítačové) nejdelší tradici. Metody matematické (zejména kvantitativní) se používaly při studiu jazykových jevů již dávno, i když se výrazu matematická (kvantitativní) lingvistika neužívalo. Podrobněji se prvními aplikacemi matematiky v lingvistice budeme zabývat v následující kapitole. Nyní si představíme alespoň ty nejdůležitější. Zásadní roli v historii kvantitativní lingvistiky hraje pojem *frekvence*. Zhruba od poloviny 19. století se začíná objevovat celá řada „výzkumů“ založených na tomto pojmu. Jednalo se o jednoduché počty různých jazykových jevů a zpravidla byly tyto počty vyvolány potřebami praxe (výuka cizích jazyků, vytváření těsnopisných soustav, sestavování Morseovy abecedy apod.). Koncem 19. století pak začínají vznikat první frekvenční slovníky. Autorem toho úplně prvního je německý stenograf F. W. Käding[6]. Frekvencí anglických hlásek se pravděpodobně jako první lingvista vůbec zabýval Američan W. D. Whitney (1827–1894) a bývá tak považován za předchůdce kvantitativní lingvistiky. Na možnosti využití matematických metod v lingvistice upozornil již v roce 1847 ruský matematik V. J. Bunjakovskij a polský jazykovědec Jan Baudouin de Courtenay roku 1904 zdůraznil, že v jazykovědě by bylo vhodné využít nejen matematiky elementární, ale i matematiky vyšší. V první polovině 20. století se o rozvoj kvantitativní lingvistiky zasloužili zejména ruský matematik A. A. Markov a americký lingvista G. K. Zipf. Markov vydal v roce 1913 statistickou analýzu textu veršovaného románu *Evžen Oněgin*. Na základě statistického zkoumání výskytu ruských souhlásek a samohlásek a pravděpodobnosti, s jakou po sobě následují v textu, došel k závěru, že je možné předvídat pravděpodobnost jejich výskytu („*markovův proces*“). Šlo o první důslednou aplikaci matematické statistiky v jazykovědě a toto dílo probudilo zájem o vzájemnou spolupráci lingvistů a matematiků. Lingvista G. K. Zipf pak ve dvacátých a třicátých letech 20. století zkoumal frekvenci hlásek a upozornil na některé obecně platné vztahy vyskytující se v přirozených jazycích, které jsou založeny na pojmu frekvence (tzv. *Zipfovy zákony*).

Různá nahodilá zkoumání jazykových jevů založená na pojmu frekvence se ale objevují i na české půdě. Koncepčněji se jimi zabýval Martin Hattala

[5]Toto označení uvádí např. Plath, W.; Cohen, M.: *Sur la statistique linguistique*. Extrait des Conférences de l'Institut de linguistique de l'Université de Paris IX, année 1949, Paris 1950, s. 8–16; Herdan, G.; Muller, Ch.: *Initiation à la statistique linguistique*. Paris 1968.

[6]Käding, F. W.: *Häufigkeits Wörterbuch der Deutschen Sprache*. Steglitz 1897.

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.