Blanka Sedlačíková
Matematická lingvistika (1)

Učitel matematiky, Vol. 10 (2002), No. 1, 30–36

Persistent URL: http://dml.cz/dmlcz/150475

MATEMATICKÁ LINGVISTIKA (1)

BLANKA SEDLAČÍKOVÁ

Druhá polovina 20. století se vyznačuje vznikem celé řady hraničních disciplín. Jednou z nich je i matematická lingvistika.

Matematickou lingvistikou rozumíme pomezní disciplínu stojící na rozhraní mezi matematikou a lingvistikou. Často se setkáváme také s názorem, že jde o tu část lingvistiky, která ve zkoumání jazyků uplatňuje různé matematické metody. Vznik této disciplíny klademe na přelom 50. a 60. let 20. století. Zpravidla se jako počátek uvádí rok 1957, kdy se konal VIII. mezinárodní lingvistický kongres v Oslu. Dnes se tradičně rozlišují 3 odvětví matematické lingvistiky podle toho, které matematické metody a postupy se uplatňují, a to:

1. *lingvistika kvantitativní*, která využívá kvantitativních matematických postupů, zejména statistiky. Proto se můžeme často setkat s označením *lingvistika statistická*. Kvantitativní lingvistika se snaží nahradit nepřesná konstatování o jazykových jevech údaji, které jsou podloženy číselně, a umožnit tak hlubší poznání jazykového systému. Na podkladě statistických dat lze také vytvářet různé modely (např. pravděpodobnostní) a porovnávat je se skutečností. Lingvistika kvantitativní jako jediná ze tří tradičních odvětví matematické lingvistiky (vedle lingvistiky algebraické a strojové) navazuje na jakousi tradici.
2. *lingvistika algebraická* je vlastně souhrn několika teorií a metod, které využívají takové matematické teorie jako algebra, matematická logika, teorie množin, teorie grafů, kombinatorika apod., neboť matematika není jen vědou o vztazích kvantitativních, jak se to může často navenek jevit, ale je to především vysoce abstraktní nástroj ke studiu systémů, tedy i ke studiu přirozeného jazyka.
3. *lingvistika strojová* (též *lingvistika počítačová, komputační*), která je vlastně praktickou aplikací dvou předcházejících odvětví matematické lingvistiky. V podstatě se jedná o využití moderní výpočetní techniky v lingvistice.

Cílem tohoto seriálu článků bude seznámit čtenáře s touto dle mého názoru velmi zajímavou disciplínou. První část bude věnována prehistorii tohoto oboru, v dalších třech částech pak budou představena jednotlivá odvětví matematické lingvistiky, a to lingvistika kvantitativní, algebraická a lingvistika strojová.

Prehistorie matematické lingvistiky

Jak již bylo zmíněno dříve, prehistorie tohoto vědního oboru se vztahuje vlastně pouze ke kvantitativní lingvistice.

První stopy užití kvantitativních metod můžeme najít již u starých Hindů. Ti počítali z náboženských důvodů slova v textu posvátných Rgvéd.

Další aplikace se pohybovaly víceméně na úrovni mystiky nebo hříčky. Můžeme sem zařadit středověké obrazové básně (carmen figuratum), různé kaligramy od Rabelaisova kaligramu *Božská láska* z Gargantuy a Pantagruela až po třeba Apollinaira. Dále lze uvést například český poetismus (sbírka *Na vlnách TSF* od Jaroslava Seiferta, báseň *Adé* ve sbírce *Pantomima* od Vítězslava Nezvala aj.), lettrismus Isidora Isoua či některé formy tzv. vizuální poezie tvořené dnešními programátory.

Přelom ve vztahu matematika versus lingvistika pak nastává s využíváním pojmu frekvence (četnost), jednoho z nejdůležitějších pojmů kvantitativní lingvistiky. Frekvence vyjadřuje počet určitého jevu v celku. V některých oborech pracujících s jazykem si odborníci všimli, že ne všechny jazykové jednotky mají stejnou frekvenci. Všechny tyto aplikace pojmu četnost se však pohybovaly na úrovni pouhé deskripce, tzn. něco se počítalo. Uveďme si několik příkladů.

J. A. Komenský, který se snažil ekonomicky rozvíjet slovní zásobu žáků, využíval frekvence slov při psaní svého díla *Ianua linguarum reserata*. Na frekvenci písmen byl brán zřetel při sestavování kazet se zásobami písmen pro tiskaře, kteří pro svou práci potřebovali mnohem více těch písmen, která mají v daném jazyce vysokou frekvenci, než písmen s frekvencí nízkou. Např. v češtině by musel mít mnohem větší zásobu písmene „a“ než písmene „f“. Podobně měla frekvence vliv na sestavování

těsnopisných systémů. Frekvenci zohledňoval také Samuel Morse, když vytvářel svoji Morseovu abecedu. Nejčastější písmena měla nejkratší kód (např. písmenu E odpovídá jedna tečka, písmenu T jedna čárka).

I na klaviatuře psacího stroje jsou nejčastější písmena na místech nejsnadněji dostupných nejobratnějšími prsty, tj. ukazováčkem a prostředníkem. Obě tyto aplikace, tzn. Morseova abeceda a klaviatura psacího stroje, jsou sestaveny optimálně pro angličtinu. I české psací stroje vychází z klaviatury anglické, jsou pouze doplněny o písmena v angličtině se nevyskytující a je provedena záměna „z" a „y". Polská matematická lingvistika v 60. letech 20. století provedla optimalizaci polských psacích strojů.

Frekvence písmen se využívala také při dešifraci textu. Uveďme si jeden zajímavý příklad z naší literární historie. Některé části svých intimních deníkových záznamů šifroval K. H. Mácha, a to tak, že si sestavil vlastní abecedu, většinou odvozenou z řeckých písmen, a každé české písmeno zaměnil symbolem z takto vytvořené abecedy. Situaci Mácha sice ještě poněkud ztížil tím, že kombinoval s češtinou německá slova a věty a každý druhý řádek (s 1 výjimkou) psal pozpátku, ale při dešifraci pak v podstatě stačilo sestavit tabulku výskytů Máchových symbolů a porovnat ji se stejnou tabulkou pro českou abecedu.

Jako první na nedostatky pouhé deskripce a na vhodnost využívání vyšší matematiky ve filologii (dříve se tímto termínem označovala lingvistika společně s literární vědou) upozornil roku 1847 v časopise *Sovremennik* V. Ja. Bunjakovskij. Vyžadoval vlastně hodnocení získaného výsledku v rámci možných hypotéz o příčině pozorovaného jevu. [7]

V této souvislosti si zaslouží pozornost neprávem zapomenutá práce našeho matematika Augustina Seydlera z roku 1886 *Počet pravděpodobnosti v přítomném sporu* opublikovaná v časopise Atheneum, kterou se zapojil do bojů o pravost Rukopisů. Připomeňme si stručně, o jaký problém se jednalo. V letech 1816 – 1818 byly objeveny dva údajně staročeské rukopisy, podle místa nálezu označované jako *Rukopis královédvorský* (RKK) a *Rukopis zelenohorský* (RKZ). Texty se vztahovaly k hrdinné české minulosti –

RKK ke 13. století, RKZ dokonce až k 10. století. Oba texty měly značný ohlas u nás i ve světě podporovaný obrozeneckým kultem minulosti. Vždyť tyto texty dokládaly starobylost české literární tradice a posouvaly ji naroveň s takovým kulturám jako byla například kultura francouzská, německá, španělská a ruská. Mladí obrozenci nepochybovali o pravosti obou památek, brzy se však ozvaly hlasy opačné. Proti RKZ vystoupila celá řada významných vědců (Jan Gebauer, Tomáš Masaryk, Jaroslav Goll aj.). Augustin Seydler pak matematickými prostředky potvrdil právě Gebauerova zjištění o nepravosti těchto rukopisů. Seydlerovi se podařilo překonat dosavadní deskriptivismus, neboť si uvědomil, která z věd je v této otázce hlavní a že řeší problém lingvistický, ne matematický, a proto je nutno získané výsledky interpretovat v rámci lingvistiky.

Nechci dokazovati matematikou něco, co může dokázati jen lingvistika, paleografie nebo historie (sociologie); vida však, jak málo se cítí váha pochybností těmito vědami, zejména první z nich pronešených, chtěl jsem, pokud to bylo lze, váhu tu číselně vyjádřiti. [7]

V 80. letech 19. století přináší první výsledky stylometrie, což je metoda vypracovaná k řešení platónských otázek , tzn. jednak k určení, které z Dialogů napsal skutečně Platón, jednak k stanovení relativní chronologie jeho děl. Zjišťovala se frekvence slov nesouvisejících s tématem, tj. hiátů, neplnovýznamových slov apod., a ta se srovnávala s frekvencí těchto slov v pravých Platónových textech. Našel-li se v jeho nesporných textech jev, jehož číselný výskyt byl zhruba stejný, a nenabýval-li této hodnoty jev v textu sporném, pak Platón není autorem tohoto textu. Dále se hledal rovnoměrný pokles či vzestup určitého jevu k určení chronologie Platónových textů. Tato metoda vychází z předpokladu existence individuálního autorského stylu podobně jako současné aplikace matematických metod při určování autorství anonymních textů.

V souvislosti s pojmem frekvence začala vznikat samozřejmě celá řada frekvenčních a konkordančních slovníků. Frekvenční slovník je vlastně seznam slov určitého celku, zpravidla v základním tvaru (u substantiv, adjektiv, zájmen a číslovek 1. pád singu-

láru, u sloves infinitiv), která jsou uspořádána buď podle frekvence nebo abecedně. Je-li slovo doplněno také informací, na jakém místě v textu se nachází, mluvíme o tzv. konkordanci. První frekvenční slovníky se objevují ke konci minulého století a vznikají zpravidla pro dobové praktické účely jako byla například konstrukce ekonomického těsnopisného systému (F. Käding, 1897), zlepšení pravopisného systému (L. Ayers, 1915), tvorba metod slepeckého čtení (J. Knowles, 1904), efektivnost vyučování cizím jazykům (R. C. Eldridge, 1911) apod. Zpočátku se zpracovávaly ručně pomocí excerpčních lístků. Aby tyty slovníky byly dostatečně vypovídající, je třeba zpracovat poměrně rozsáhlý materiál, což nastává až s rozvojem počítačů. Proto si o frekvenčních a konkordančních slovnících povíme více později.

Výraznou roli v rozvoji kvantitativní lingvistiky sehráli v první polovině 20. století dva vědci, A. A. Markov a G. K. Zipf. V roce 1913 vydává ruský matematik A. A. Markov práci *Příklad statistického výzkumu textu Evžena Oněgina* ... (Primer statističeskoho issledovania nad tekstom "Evgenija Onegina" illjustrirujuščij svjazď ispytanij v cep'). Dochází zde k závěru, že v každé části textu lze s určitou pravděpodobností předpokládat, které jazykové jednotky budou dále následovat. Byl tak popsán jev, který známe z teorie pravděpodobnosti pod názvem *Markovův proces.* Jeho podstata je následující: mluvení je proces, ve kterém k již existujícím jednotkám (vysloveným či napsaným) přiřazujeme jednotky nové, a to na základě jejich relativní frekvence, která je v daném jazyce závazná. Ukažme si jeden zajímavý příklad s nápodobou českého, anglického a německého textu podle teorie pravděpodobnosti. V češtině existuje 42 různých písmen, počítáme-li také mezeru a nerozlišujeme-li písmena „ů" a „ú". Nejprve budeme předpokládat stejnou frekvenci u všech písmen abecedy, dále budeme přihlížet k relativní frekvenci jednotlivých písmen, dvojic písmen a trojic písmen. Dostaneme následující výsledky:

1) Za předpokladu, že všechna písmena v textu mají stejnou frekvenci:
čeština: ďj mrgučxýďyaýweaožá
angličtina: xfoml rxkhrjff juj zlpwcfwkkcyj

němčina: aiobnin tarsfneoulpiitdregedcoads

2) S přihlédnutím k relativní frekvenci jednotlivých písmen:
čeština: žia ep atndi zéuořmp
angličtina: ocro hli rgwr nmielwis eu ll
němčina: er agepterprteiningeit gerelen re

3) S přihlédnutím k relativní frekvenci dvojic písmen:
čeština: lí di oneprá sguluvicechupsv
angličtina: on ie antsoutinys are t inctore
němčina: billunten zugen hin se sch wel

4) S přihlédnutím k relativní frekvenci trojic písmen:
čeština: dves a vaše miléklár
angličtina: in no ist lat whey cratict froure
němčina: eist des nich in den plassen kann

Vidíme, že pokud uvažujeme u všech písmen stejnou frekvenci jejich výskytu, dostáváme text, který prakticky vůbec nepřipomíná text příslušného jazyka. Uvažujeme-li však relativní frekvenci nějaké n-tice písmen, dostáváme při rostoucím přirozeném čísle n text, který se přibližuje textu daného jazyka. Pokusy v některých jazycích ukázaly, že již při $n = 32$ se takto vzniklý text prakticky rovná textu skutečného jazyka.

Významně k rozvoji kvantitativní lingvistiky přispěl rovněž americký lingvista německého původu George Kigsley Zipf. Ten se na přelomu 20. a 30. let zabýval frekvencí hlásek a došel k několika zajímavým poznatkům (např. čím je hláska artikulačně obtížnější, tím menší je její frekvence; ve všech jazycích jsou neznělé hlásky přibližně dvakrát častější než znělé aj.). Tři nejdůležitější zákonitosti jsou známy jako tzv. Zipfovy zákony:

První Zipfův zákon: *Součin frekvence slova a jeho ranku je konstantní.* (Rank je pořadí slova v seznamu podle klesající frekvence.)

Druhý Zipfův zákon: *Počet slov o jisté frekvenci krát frekvence na druhou je konstantní.*

Třetí Zipfův zákon: *Slova s vysokou frekvencí mají zpravidla větší počet významů.*

A na tomto místě můžeme uzavřít prehistorii matematické lingvistiky a dále se věnovat pouze historii.

Mgr. Blanka Sedlačíková
doktorandka Katedry matematiky PřF MU
Janáčkovo nám. 2a, 662 95 Brno
email: hvezdova@math.muni.cz