Blanka Sedlačíková
Matematická lingvistika (2)

Učitel matematiky, Vol. 10 (2002), No. 2, 80–88

Persistent URL: http://dml.cz/dmlcz/150485

MATEMATICKÁ LINGVISTIKA (2)

BLANKA SEDLAČÍKOVÁ

Kvantitativní lingvistika

Termínem *kvantitativní lingvistika* označujeme jedno ze tří dnes rozlišovaných odvětví matematické lingvistiky (vedle lingvistiky algebraické a strojové), které využívá kvantitativních matematických metod, tj. statistiky, počtu pravděpodobnosti, matematické statistiky, teorie informace apod. Nejčastěji jsou uplatňovány postupy statistické, proto také někdy hovoříme o *statistické lingvistice*. Cílem kvantitativní lingvistiky je lepší poznání jazykového systému tak, že vágní informace o různých jazykových jevech (hláska, písmeno, foném, slabika, slovo, pád, čas, typ věty aj.) jsou nahrazeny údaji podloženými číselně. Na základě těchto údajů lze vytvářet modely, popřípadě je porovnávat se skutečností. Kvantitativní lingvistika navazuje na poměrně dlouhou tradici, o které jsme si již něco řekli minule. Její počátek však klademe až na konec 50. let minulého století, kdy vznikla teorie informace, která ji do jisté míry ovlivnila.

Kvantitativní lingvistika se zpravidla dále dělí podle toho, které jazykové jevy a jejich vztahy jsou předmětem jejího studia. Slovo a jeho postavení ve slovní zásobě je předmětem *statistiky lexikální*, tvar slova a jeho morfologické kategorie jsou předmětem *statistiky morfologické* a věta, syntagma a syntaktické kategorie pak *statistiky syntaktické*. Jde o tři základní oblasti kvantitativní lingvistiky. Dále můžeme rozlišit například *statistiku fonologickou*, *sémantickou*, *grafematickou*, *slovotvornou*, *morfematickou*, *stylistickou*, *typologickou* či *nářeční*. Kvantitativní metody nacházejí své uplatnění také v *textologii* (jde zejména o problematiku sporného autorství) či v oblasti *vývoje jazyka*.

Nejpropracovanější oblastí kvantitativní lingvistiky je *lexikální statistika*, jejímž úkolem je kvantifikovat slovní zásobu a její složky.

Jedním z nejvýznamnějších problémů, kterými se zabývá, je tvorba **frekvenčních a konkordančních slovníků**. Frekvenčním slovníkem (dále FS) rozumíme seznam slov určitého souboru uspořádaný buď podle frekvence nebo abecedně s údajem o frekvenci. Konkordanční slovník je vlastně FS doplněný navíc o informaci, kde se jednotlivá slova v textu nacházejí. Tvorba těchto FS má již více než stoletou historii a jejich přínos pro jazykovědu je značný. Obsahují totiž řadu užitečných údajů o slovní zásobě i o jejím vztahu k rovině gramatické či sémantické. Zajímavé mohou být však také pro matematiky, informatiky, psychology, filozofy, pedagogy, metodiky, stenografy, kryptology apod. Jednotlivé FS se mohou lišit nejen jazykem, pro který vznikají, ale i rozsahem zpracovávaného materiálu, výběrem textů či způsobem zpracování (ručně, pomocí počítačů). Proto dnes některé jazyky mají již celou řadu FS, zatímco jiné pouze jeden nebo dokonce nemají žádný.

Za první FS považujeme *Häufigkeitswörterbuch der deutschen Sprache* („Slovník četnosti výskytu německého jazyka") německého stenografa F. W. Kädinga z let 1897-98. Vedle svého prvenství je významný i obrovským rozsahem materiálu (10 910 777 slov), proto se k němu později vracela celá řada lingvistů a pedagogů. Za zmínku stojí i další německé FS–slovníky odborných disciplín (medicína, chemie, fyzika) zpracované jako slovníky dvojjazyčné (rusko-, anglicko-, francouzsko-německé), FS západoněmeckých novin (obsahuje cca 6 500 000 slov z Die Welt a 6 000 000 slov z Süddeutsche Zeitung z let 1967-68) či FS hovorové horní němčiny, který polovinu materiálu čerpá z denního tisku a zábavných časopisů a druhou polovinu z magnetofonových nahrávek mluveného jazyka.

Jedním z nejstarších anglických FS je slovník L. P. Ayrense z roku 1915, který zpracoval 368 000 slov z obchodních a soukromých dopisů. Ojedinělý je sémantický FS M. Westa z roku 1973, v němž udává frekvenci různých významů slov.

Mezi FS románských jazyků jsou nejzajímavější slovníky Al. Juillanda, který s různými spoluautory vydal FS španělštiny, rumunštiny, francouzštiny, portugalštiny a italštiny. U všech materiál

pochází z pěti žánrů (drama, umělecká próza, eseje, technická literatura a periodika), celkový rozsah je půl miliónu slov (vždy po 100 000 slov z každého žánru), vedle frekvenčního seznamu uvádí i abecední seznam slov s frekvencí alespoň 3 a všechny tyto FS jsou zpracovány počítačově. Vedle přínosu teoretického (např. zavedení koeficientu disperze a užití slova) jsou tyto FS významné tím, že díky stejným kritériím výběru materiálu a stejnému zpracování umožňují srovnání různých románských jazyků. Rozsahem materiálu je zajímavý slovník E. Bruneta z roku 1981, který zachycuje vývoj francouzštiny od roku 1789 do současnosti a který zahrnuje texty o délce 70 273 552 slov. FS ruštiny z roku 1977 vytvořený pod vedením L. N. Zasorinové zkoumá materiál kolem jednoho miliónu slov a rovnoměrně ho vybírá ze čtyř skupin – beletrie, drama, věda, publicistika. Podle něj zaujímá prvních 9 044 nejfrekventovanějších slov 92,4% textu a dalších asi 30 000 potom zbývajících 7,6%. Podle FS, který zpracovává jazyk A. S. Puškina, jeho slovník tvoří 21 197 různých slov a prvních 2 000 nejfrekventovanějších pokrývá 80% textů. Za zmínku stojí rovněž „Slovník asociačních norem" z roku 1977, jehož materiál byl získán tak, že na 556 slov odpovídalo 200-700 informátorů prvním slovem, které je napadlo. Získaná slova pak slouží při výuce cizinců, neboť jde vlastně o nejdůležitější frazeologismy či syntagmatická a paradigmatická spojení slov.

FS češtiny z roku 1961 nese název *Frekvence slov, slovních druhů a tvarů v českém jazyce* (zkratka FSČ) a jeho autory jsou J. Jelínek, J. V. Bečka a M. Těšitelová. Autoři ručně zpracovali 75 textů z 8 stylistických oblastí (beletrie, poezie, literatura pro mládež, drama, odborná literatura, žurnalistika, vědecká literatura a mluvené rozhlasové projevy) a celkem tak získali 1 623 527 slov. U každého slova v seznamu prvních 10 000 nejfrekventovanějších je vedle celkové frekvence uveden i počet stylistických oblastí a děl, ve kterých se dané slovo vyskytlo, např. *Blažena* 492-2-02 nebo *platiti* 365-8-66. Ačkoliv je frekvence slova *Blažena* vyšší, vyskytuje se pouze ve dvou stylistických oblastech a ve dvou dílech. Jde tedy o jev náhodný a užití tohoto slova ve skutečném jazyce bude nižší než slova platiti, neboť to se vyskytlo ve všech 8 stylistických

oblastech a v 66 dílech ze 75. Vedle toho FSČ obsahuje abecední seznam sestávající z 26 257 nejfrekventovanějších slov až do frekvence 3, jenž vedle absolutní frekvence uvádí i počet stylových skupin a počet textů (s distribucí v rámci jednotlivých skupin), v nichž se dané slovo vyskytuje. Slovník je také doplněn o teoretickou část, ve které jsou shrnuty dosavadní poznatky z oblasti kvantitativní lingvistiky. Bohužel dnes je tento FS již zastaralý a ačkoliv pro češtinu vznikla celá řada dílčích FS (např. publicistiky, administrativy, odborné češtiny, věcného stylu), je úkolem dnešní kvantitativní lingvistiky sestavení nového FS, který by odpovídal aktuálnímu stavu jazyka.

Frekvenční slovníky, ale i jiné práce z oblasti kvantitativní lingvistiky přinesly celou řadu nových poznatků o přirozených jazycích. Byla zavedena také celá řada pojmů, které slouží ke srovnávání jednotlivých jazyků, stylů, k určování sporného autorství apod. Ukažme si pro představu alespoň pár z nich právě na příkladu slovní zásoby.

Nejprve si uveďme prvních deset nejfrekventovanějších slov podle FSČ. Jsou to: *a, být, ten, v(e), on, na, že, s(e), z(e), který.* Těchto deset slov řadíme do **pásma slov s nejvyšší frekvencí.** Jedná se o slova velmi krátká, zpravidla jednoslabičná (výjimku zde tvoří pouze zájmeno *který*, jež však bývá v hovorové češtině nahrazováno jednoslabičným *co*), což souvisí s projevem jisté ekonomie jazyka (tzn. nejčastěji používaná slova jsou co nejkratší a artikulačně co nejjednodušší, aby byla námaha mluvidel minimální). Tato slova pokrývají v průměru asi 20% textu (na první nejčastější připadne asi 5% textu, na desáté 1%). Znalost těchto kvantitativních vztahů ve slovní zásobě je spolu se znalostí frekvence jednotlivých písmen důležitá při dešifrování tajných kódů. Vidíme, že nejčastější slova v češtině jsou slova formální (tzn. předložky, spojky, pomocná slovesa apod.), neboť jejich počet je značně omezený. A jak je tomu v ostatních jazycích? Například ve slovenštině jsou to slova: *a, byť, v, na, sa, ten, on, že, z, jako*, v ruštině: *v(o), i, ne, na, ja, byť, čto, on, s(o), a*, v angličtině: *the, of, and, a, in, that, is, was, he, for.* Stejně jako v češtině ve všech těchto jazycích s ní příbuzných i nepříbuzných jsou nejčas-

tější slova formální a velmi krátká. Pravděpodobně jde o vlastnost společnou všem indoevropským jazykům. Do **pásma slov se střední frekvencí** řadíme slova od pořadí 11 do frekvence 11 a jeho rozsah závisí na velikosti korpusu, stylu apod. Třetí pásmo, **slova s frekvencí nižší a nejnižší**, tvoří velké množství jednotek s frekvencí 10 až 1. Toto pásmo má vliv na tzv. **bohatství slovníku**, zatímco pásmo slov s nejvyšší frekvencí má vliv na tzv. **koncentraci slovníku**. Jak jsme již uvedli, FSČ zpracovává celkem 75 textů a všechna slova těchto textů tvoří tzv. **délku textu** (značíme N) – zde 1 623 527 slov. Některá slova se v textech s větší či menší frekvencí opakují, proto je v těchto textech mnohem méně slov různých. Tato různá slova tvoří **slovník** (značíme V) – ve FSČ je 54 486 slov různých. Mezi délkou textu N a slovníkem V existují různé vztahy. Například součet frekvencí jednotlivých částí V se rovná N. S rostoucím N roste i V, ne však proporcionálně (v románě E. Basse *Lidé z maringotek* o délce textu 47 542 slov je slovník 8 673 různých slov, v románě V. Řezáče *Černé světlo* o 55 164 slovech je slovník takřka stejně velký, a to 8 675 slov). Pro výpočet bohatství slovníku (značíme R) pak podle P. Guirauda platí:

$$R = \frac{V}{\sqrt{N}} \qquad \text{(platí pro všechna slova)}$$

$$R = \frac{V}{\sqrt{2N}} \qquad \text{(platí pro plnovýznamová slova)}$$

Koncentraci slovníku (C) vypočítáme takto:

$$C = \frac{\sum_{50}^{1}}{N}$$

(poměr prvních padesáti nejfrekventovanějších plnovýznamových slov k délce textu).

Marie Těšitelová, která je jedním z nejvýznamnějších vědců zabývajících se otázkami kvantitativní lingvistiky u nás, zjistila, že oba tyto vzorce pro češtinu neplatí. V případě bohatství slovníku totiž Guiraud předpokládá, že poměr plnovýznamových a

neplnovýznamových slov je ve francouzštině 1 : 1, zatímco v češtině je tento poměr podle Těšitelové 4 : 1. Rozdíl je způsoben pravděpodobně tím, že francouzština jako analytický jazyk obsahuje velké množství formálních (neplnovýznamových) slov oproti češtině, která je jazykem flexívním, a tedy vyjadřuje různé kategorie pomocí koncovek. Koncentrace slovníku je podle Guirauda konstantní, neboť prvních padesát nejfrekventovanějších plnovýznamových slov ve všech textech představuje 18%. M. Těšitelová však zjistila, že pro češtinu se tato hodnota pro různé texty pohybuje v rozmezí 13,35 – 25,05%. Znázorníme-li si frekvenční seznam graficky, dostaneme křivku exponenciálního rozložení.

Velmi zajímavé jsou odhady slovní zásoby jednotlivých jazyků, individuální slovní zásoby apod.:

1. **Průměrný slovník dětí** (odhady se u různých autorů liší, což může být způsobeno nerozlišováním slovní zásoby aktivní a pasivní):
   1 rok - 10 slov
   2 roky - 300 slov
   3 roky - 900 slov
   4 roky - 1 650 slov
   5 let - 2 500 slov
   6 let - 3 500 slov
   14 let - 9 000 - 19 500 slov

2. **Slovník dospělých podle profesí** (Ogden):
   farmář - 300 slov
   japonský diplomat - 7 000 slov
   univerzitní student - 12 000 slov
   James Joyce - 250 000 slov

3. **K běžnému dorozumění** v cizím jazyce je podle Ogdena třeba znát asi 850 slov (600 podstatných jmen, 150 přídavných jmen, 100 sloves). K sledování odborné literatury potom ještě dalších 150 (100 termínů obecně vědeckých a 50 termínů z daného oboru) – celkem tedy 1 000 slov.

4. **Slovní zásoba některých slovníků**:
   Příruční slovník jazyka českého - asi 250 000 slov

Slovník spisovného jazyka českého - 192 908 slov
Slovník spisovné češtiny - 47 559 slov
Slovník slovenského jazyka - asi 120 000 slov
Ruský slovník Dálův - asi 200 000 slov
Francouzský slovník Littréův - asi 210 000 slov
(tzn. průměrný slovník současného spisovného jazyka je přibližně 200 000 slov)

5. **Slovní zásoba některých spisovatelů:**
A. France - 9 000 slov
Homér - 9 000 slov
J. W. Goethe - 20 000 slov
A. S. Puškin - 21 200 slov
W. Shakespeare - 24 000 slov

6. **Slovník některých děl:**
K. Čapek: Obyčejný život - 5 539 slov
K. Čapek: Život a dílo skladatele Foltýna - 4 145 slov
M. Pujmanová: Předtucha - 4 858 slov
Fr. Halas: Ladění - 2 078 slov
J. Hora: Kniha domova - 2 961 slov
odborný text (při průměrné délce 18 000-23 000 slov) - 3 000 - 3 500 slov
publicistický text (při rozsahu cca 19 000 slov)-4 700 slov

Kvantitativní lingvistika se samozřejmě zabývá také frekvencemi na úrovni ostatních rovin jazyka. Uveďme si zde několik příkladů na *úrovni fonetické*. Čeština má celkem 36 různých hlásek. Prvních 9 nejfrekventovanějších vypadá takto: e (9,79%), o (6,91%), a (6,66%), i (6,00%), t (4,76%), s (4,71%), n (4,56%), l (4,35%), k (4,02%). Poměr samohlásek a souhlásek je 41 : 59, poměr krátkých a dlouhých samohlásek je 78 : 22 a poměr znělých souhlásek k neznělým je 63 : 37. Co se týče dvoučlenných kombinací českých hlásek, tak z celkového možného počtu 1296 kombinací se jich realizuje kolem 60% a nejčastější jsou tyto: je, st, ne, na, po, se, /ňí/, ro, /ňe/, en, le, em, la, ov, li, to, ko, te, el, pr. Slova v českém textu začínají nejčastěji hláskami s, p, n, v a první samohláska a se objevuje až na 9. místě. Souhláskou začíná

celkem 88% slov a samohláskou 12% slov. Na konci slov se potom nejčastěji objevují hlásky e, i, a, í, o. Samohláskou končí 71% slov a souhláskou 29%.

Na závěr si ještě uveďme jednu zajímavou metodu zabývající se časovým určováním *vzniku jazyka*. Tato metoda nazývaná *glottochronologie* (doslova vlastně chronologie jazyka) nebo také *lexikostatistika* se objevuje v 50. letech 20. století. Její rozvoj je spjat se dvěma americkými badateli, a to Morrisem Swadeshem a Robertem B. Leesem. Cílem glottochronologie je pomocí kvantitativních metod zjistit, ve které době došlo k rozrůznění určitého jazyka nebo prajazyka na dva nebo více jazyků moderních. Příbuzenské vztahy jazyků se měří na základě změn v slovní zásobě, lexiku (odtud název lexikostatistika). Metoda byla inspirována tzv. rozpadovým zákonem, na jehož základě lze určit stáří organických látek. Vycházelo se ze 2 předpokladů: 1) každý jazyk má tzv. jádro slovní zásoby, tj. několik desítek do značné míry stabilních výrazů, které označují základní věci v životě člověka (např.: *matka, otec, pták, ryba, pes, muž, žena, dlouhý, velký, malý, já, ty, my, tento, onen, kdo, co, všechno, mnoho, jeden, dva, ne, jíst, pít, kousat, vidět, slyšet, znát, spát* apod.); 2) v tomto jádru dochází ke změnám poměrně pomalu, ale hlavně s konstantní rychlostí. Při zjišťování časové hloubky, tzn. období, kdy došlo k rozrůznění příslušných jazyků, postupujeme tak, že stanovíme asi 100 výrazů základního jádra slovní zásoby v těchto jazycích, jejich porovnáním zjistíme procento shodných a různých dvojic a tzv. index rychlosti mizení slov z jádra. Vzorec pro výpočet časové hloubky vypadá potom takto:

$$i(t) = \frac{\log C}{2 \log r}$$

$i(t)$ = časová hloubka (uběhlý čas)
C = procento shodných dvojic
r = index rychlosti mizení slov z jádra (procento dvojic uchovaných za určitou jednotku času)

Tato metoda získala celou řadu zastánců. U nás ji pro zjištění doby rozpadu praslovanštiny na větev západní, východní a jižní použili brněnští jazykovědci M. Čejka a A. Lamprecht. Užitím

této metody bylo zjištěno, že rozpad praslovanštiny nastal v 8.-11. století. Objevila se samozřejmě i celá řada kritiků (E. Coseriu, W. W. Arnold). Bylo jí vytýkáno to, že výběr slov jádra a jejich počet je subjektivní a že i rychlost rozpadu není vždy konstantní, neboť je ovlivňována vnějšími vlivy.

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

★ ★ ★

OZNÁMENÍ

Komise pro vzdělávání učitelů matematiky a fyziky při ÚV JČMF pořádá ve dnech 19. – 22. srpna 2002 na Gymnáziu v Jevíčku

XI. SEMINÁŘ O FILOZOFICKÝCH OTÁZKÁCH MATEMATIKY A FYZIKY

Předběžné finanční náklady: vložné 250 Kč, nocležné 100 Kč za noc, stravné 100 Kč na den. Přihlášku a podrobnější informace je možno získat na seminární adrese:

RNDr. Aleš Trojánek
Gymnázium, Velké Meziříčí, Sokolovská 27
594 01 Velké Meziříčí
tel.,fax: 0619 521 600
e-mail: trojanek@gvm.cz
http://www.gvm.cz