2.9 Glottochronologie

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (Czech). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 101–104.

Persistent URL: http://dml.cz/dmlcz/402324

jakákoliv chyba při psaní či tisku mohla zcela změnit význam sdělení. Skutečnost je ale zcela jiná. Má-li čeština 42 písmen, z toho 14 „*samohláskových*“ a 28 „*souhláskových*“, pak by při nejekonomičtějším využití české abecedy existovalo 392 slov tvaru „*samohláska + souhláska*“ (tj. $14 \cdot 28 = 392$) a 42 slov tvořených pouze jedním písmenem. Ve skutečnosti jen malá část z nich jsou česká slova – dvoupísmenná například *ač*, *ach*, *ať*, *au*, *as*, *až*, *ob*, *oč*, *od*, *och*, jednopísmenná *a*, *i*, *o*, *u*, *k*, *s*, *v*, *z* (jednopísmenných je jen zhruba $\frac{1}{4}$). A pouze díky redundanci pak můžeme správně porozumět i zkomolenému textu:

PRIJEDU ZIBRA VEKER.

Hodnoty příslušné redundance zjištěné z hodnot entropie pro různé jazyky kolísají v rozmezí $0{,}70 \pm 0{,}10$. Nelze říci, jestli je toto kolísání dáno skutečnými odchylkami redundance v různých jazycích nebo jestli je to následek nejednotnosti metody v zjišťování redundance. V každém případě se tu naskýtá rozsáhlá oblast jazykovědných výzkumů. Pokud se opravdu hodnoty redundance v různých jazycích různí, pak je třeba zkoumat, jakými vlastnostmi konkrétního jazyka je hodnota redundance podmíněna. Pokud se ukáže, že hodnoty redundance jsou pro všechny jazyky zhruba stejné, pak nás napadá řada dalších otázek:

1) Proč má redundance právě tuto zjištěnou hodnotu? Zde pravděpodobně může být nápomocna psychologie, fyziologie apod., neboť tato hodnota bude nejspíše ovlivněna podmínkami mluvení, vnímání řeči atd.

2) Jaké je místo oněch prostředků, postupů, kterými se v přirozených jazycích dosahuje zjištěné hodnoty redundance, mezi všemi takovými postupy? Zde by mohlo být nápomocno srovnávací studium přirozených jazyků a umělých kódů studovaných a konstruovaných v teorii kódování.

3) Proč se v přirozených jazycích uplatňují právě tyto postupy?

Tyto výzkumy jsou velmi cenné, neboť začleňují tyto ukazatele do jiných poznatků o jazyce. A rovněž svůj praktický význam měly některé jednoduché aplikace teorie informace na studium přirozených jazyků – na základě jejich poznatků byly například sestavovány telegrafní kódy novoindických jazyků.

2.9 Glottochronologie

Zajímavou aplikací matematiky v jazykovědě je tzv. *glottochronologie* (též *lexikostatistika*). Je to lexikologická metoda, která pomocí statistiky zjišťuje dobu vzniku jazyka, respektive různých jazyků. Tato metoda vznikla v 50. letech 20. století a za jejího zakladatele je považován americký antropolog Morris Swadesh[152]. Ve stejné době uvádí podobné výsledky i Američan Robert B. Lees[153].

[152] Swadesh, M.: *Lexico-Statistic Dating of Prehistoric Ethnics Contacts.* Proceedings of the American Philosophical Society 96, 1952, 452–463.

[153] Lees, R. B.: *The Basis of Glottochronology.* Language 29, 1953, 113–127.

Obrázek 2.2: Morris Swadesh

Cílem glottochronologie je pomocí kvantitativních metod zjistit, v 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*).

Glottochronologie byla inspirována tzv. *radiokarbonovou metodou*, u jejíhož vzniku stál americký chemik Willard Frank Libby a pomocí níž lze stanovit stáří organických látek. Po zániku organických látek dochází k rozpadu radioaktivních jader uhlíku a tento rozpad se řídí exponenciální funkcí, to znamená, že po uplynutí určitého časového intervalu klesne počet radioaktivních jader na polovinu, po uplynutí téhož času znovu klesne na polovinu atd. Tento časový interval, během něhož klesne počet radioaktivních jader uhlíku na polovinu, se nazývá *poločas rozpadu*. Pro izotop uhlíku stanovil Libby poločas rozpadu na 5 568 let. V roce 1960 byla Libbymu udělena za tuto metodu Nobelova cena za chemii. Později se přišlo na to, že poločas rozpadu izotopu uhlíku byl stanoven nepřesně a pomocí tzv. *dendrochronologické kalibrace* byla jeho hodnota upřesněna na 5 730 let.

A právě tato radiokarbonová metoda inspirovala na přelomu 40. a 50. let 20. století M. Swadeshe k její aplikaci na vývoj jazyků. V první verzi glottochronologie z roku 1952 určil jakési *jádro slovní zásoby*, nejprve o 215, později o 200 lexikálních jednotkách, u nichž předpokládal, že jsou dostatečně univerzální a stabilní, nezávislé na kulturních či zeměpisných podmínkách (např. *matka*, *otec*, *pták*, *ryba*, *pes*, *muž*, *žena*, *dlouhý*, *velký*, *malý*, *znát*, *spát* apod.). Předpokládal, že v tomto jádru slovní zásoby se výměna slov řídí obdobným exponenciálním zákonem jako rozpad radioaktivních jader izotopu uhlíku, tj. ke změnám dochází poměrně pomalu a mizení slov ze slovníkového jádra je konstantní. Ve spolupráci se specialisty pak zjišťoval, jakou rychlostí dochází ke změnám v tomto jádru, a to u jazyků s dostatečně dlouhou literární tradicí. Průměrná hodnota změn v testovacím souboru se pohybovala okolo 19,5 % za časový interval tisíc let. Po kritikách glottochronologie publikoval Swadesh v roce 1955 novou studii, v níž slovníkové jádro redukoval na 100 jednotek, některé pojmy slovníkového jádra nahradil pojmy univerzálnějšími a přepočítal i *rozpadovou konstantu* λ, jejíž hodnotu stanovil na 14 %, to znamená, že za tisíc let zůstane zachováno 86 % slov ze slovníkového jádra.

Nejčastěji se při zjišťování *časové hloubky* (tzn. období, kdy došlo k rozrůznění zkoumaných jazyků) používají tyto dva vzorce:

$$i_{(t)} = \frac{\ln C}{2 \ln r} \tag{2.1}$$

$$i_{(t)} = \frac{\ln C}{-2\lambda} \tag{2.2}$$

kde
$i_{(t)}$ ... časová hloubka (uběhlý čas);
C ... procento společně zděděných párů slov z počtu všech párů slov v slovníkovém jádru obou zkoumaných jazyků;
r ... index rychlosti mizení slov z jádra (procento dvojic uchovaných za určitou jednotku času, zpravidla tisíc let), index r je komplementární k rozpadové konstantě λ (tzn. $r = 1 - \lambda$), dle Swadeshových výpočtů klademe $r = 0{,}86$;
λ ... rozpadová konstanta (konstanta komplementární k r, tzn. $\lambda = 1 - r$), $\lambda = 0,14$.

Při zjišťování *časové hloubky* postupujeme tak, že stanovíme zpravidla 100 výrazů základního jádra slovní zásoby v těchto jazycích a jejich porovnáním zjistíme procento shodných a různých dvojic. Dosazením této hodnoty a indexu rychlosti mizení slov (popř. rozpadové konstanty) do příslušného vzorce získáme hledanou časovou hloubku.

P. Novotná ve své rigorózní práci [42] uvádí postup při hledání doby divergence němčiny a francouzštiny. V základním slovníkovém jádru obou jazyků nacházíme 33 párů společně zděděných slov, lze tedy položit $C = 0,33$. Dosazením do rovnice (2.1) dostaneme dobu divergence v tisíciletích:

$$i_{(t)} = \frac{\ln 0{,}33}{2 \ln 0{,}86} = \frac{-1{,}10886}{-0{,}30164} = 3{,}675.$$

Jiná situace nastává v případě, že nemáme synchronní dvojici příbuzných jazyků, tzn. každý z dvojice studovaných jazyků je zachycen v jiné době (od rozpadu jejich společného prajazyka do doby jejich zachycení uběhla různě dlouhá časová období $i_{(t_1)}$ a $i_{(t_2)}$). V tomto případě Swadesh vychází ze vztahu (2.1), který lze přepsat jako

$$2i_{(t)} = \frac{\ln C}{\ln r}, \tag{2.3}$$

a ze vztahu

$$i_{(t_1)} + i_{(t_2)} = \frac{\ln C}{\ln r}. \tag{2.4}$$

Zpravidla neznáme ani jeden z časových úseků $i_{(t_1)}$ a $i_{(t_2)}$, jen jejich rozdíl $\triangle i_{(t_{12})}$. Nahradíme proto součet $i_{(t_1)} + i_{(t_2)}$ výrazem $i_{(t_1)} + i_{(t_1)} + \triangle i_{(t_{12})} = 2i_{(t_1)} + \triangle i_{(t_{12})}$, kde $i_{(t_1)}$ je kratší z obou časových úseků $i_{(t_1)}$, $i_{(t_2)}$. Odtud dostáváme pro dva asynchronně zaznamenané jazyky vztah

$$i_{(t_1)} = \frac{\ln C}{2 \ln r} - \frac{\triangle i_{(t_{12})}}{2}, \tag{2.5}$$

kde

$$i_{(t_1)} = min(i_{(t_1)}, i_{(t_2)}).$$

Glottochronologie měla pozitivní ohlas u lingvistů zabývajících se jazyky bez dlouhé literární historie, naopak kritického přijetí se jí dostalo zejména od indoevropeistů, neboť u těchto jazyků s delší literární tradicí se výsledky glottochronologických výpočtů výrazně lišily od již známých výsledků. K nejvýraznějším kritikům patřili např. E. Coseriu a W. W. Arnold. Nejčastěji bylo glottochronologické metodě vytýkáno to, že výběr slov jádra a jejich počet je značně subjektivní, což je v rozporu s přesným aparátem zpracování. Dále bylo upozorňováno na to, že rychlost rozpadu slovníkového jádra není vždy konstantní, neboť na ni mají vliv i vnější vlivy.

Zajímavé modifikace lexikostatistické metody předložili zejména Kanaďanka Sheila Embletonová a Rus Sergej Starostin. Oba autoři se shodují v tom, že Swadeshova glottochronologie nerozlišuje záměnu slova za inovaci od náhrady výpůjčkou (více viz [42]). Starostin navrhuje proto nejprve očistit základní slovníkové jádro od jakýchkoliv výpůjček. Výsledkem je menší hodnota rozpadové konstanty λ. Ve svých úvahách jde ale mnohem dále a odvozuje nové vzorce pro výpočet časové hloubky (tzv. *rekalibrovaná glottochronologie*).

Metody glottochronologie využívala ke studiu různých jazyků celá řada autorů. Ke studiu indiánských a eskymáckých jazyků ji jako první použil sám M. Swadesh. V 60. letech ji aplikoval na studium příbuznosti jazyků kavkazských G. A. Klimov. U nás se problematikou diferenciace slovanských jazyků pomocí lexikostatistické metody zabývali brněnští jazykovědci M. Čejka a A. Lamprecht[154]. S její pomocí se pokusili vymezit dobu rozpadu praslovanské jednoty na větev západní, východní a jižní. Nedostatky metody se snažili vyrovnat rozborem fonologickým a morfologickým. Rozpad praslovanštiny stanovili na konec 10. století (podle glottochronologie 8. – 11. stol.). V současnosti se glottochronologické metodě věnují na FF Masarykovy univerzity v Brně P. Novotná a V. Blažek. Bibliografii vztahující se ke glottochronologii, ale i k problematice využívání kvantitativních metod v historické a srovnávací lingvistice či dialektologii, lze nalézt rovněž u M. Těšitelové (viz [66]).

2.10 Jan Amos Komenský

Ve svém článku *J. A. Komenský a racionalizace osvojování si slovní zásoby*[155] uvádí Marie Těšitelová, že Jan Amos Komenský byl jedním z prvních autorů, který využíval frekvence slov k ekonomickému rozvíjení slovní zásoby žáků, a to ve svém díle *Janua linguarum reserata* (*Brána jazyků otevřená*). Autorka vychází z článku D. Harkina[156], ve kterém autor čtenáře seznamuje s historií měření slovní zásoby k nejrůznějším účelům – například tiskařská

[154] M. Čejka – A. Lamprecht: *K otázce vzniku a diferenciace slovanských jazyků*. Sb. fil. fak. brněnské university 1963, A 11, s. 1–20; M. Čejka: *Lexicostatistic Dating and Slavonic Languages*. Sb. fil. fak. brněnské university 1972, A 20, s. 39–52.

[155] Slovo a slovesnost 31, 1969, s. 187–191.

[156] *The History of Word Counts*. Babel (Revue Internationale de la traduction) 2, 1957, č. 3, s. 113–124.