2.6 George Kingsley Zipf

In: Blanka Sedlačíková (author): Historie matematické lingvistiky. (Czech). Brno: Akademické nakladatelství CERM v Brně, 2012. pp. 62–68.

Persistent URL: http://dml.cz/dmlcz/402321

## 2.6 George Kingsley Zipf

O kvantitativní lingvistice se někdy mluví také jako o Zipfově lingvistice podle amerického lingvisty a psychologa německého původu George Kingsleyho Zipfa z Harvardovy univerzity. Zipf zkoumal ve 20. a 30. letech 20. stol. relativní frekvenci hlásek a došel k několika zajímavým závěrům – např. čím je hláska artikulačně obtížnější, tím menší je její frekvence; ve všech jazycích jsou neznělé hlásky přibližně dvakrát častější než znělé. Rovněž se zajímal o psychologické a fyziologické faktory ovlivňující produkci a percepci řeči[89]. Hlavní zásadou je podle něj tendence k co nejmenší námaze mluvčího. Tuto svou teorii nazývá *psychobiologií* a navrhuje vyčlenit v lingvistice zvláštní disciplínu, tzv. *biolingvistiku*[90]. Největší Zipfův přínos ale spatřujeme ve třech zákonech (nazýváme je *Zipfovy zákony*), které upozorňují na vztah mezi frekvencí slov a jejich pořadím, dále na vztah mezi frekvencí slova a počtem různých slov, která tuto frekvenci mají, a na vztah mezi frekvencí slova a počtem jeho významů.

**První Zipfův zákon**:

$$r \cdot f = k$$

(součin frekvence slova a jeho ranku je konstantní), kde
$r$ ... rank (pořadí slova v seznamu podle klesající frekvence, kde jednomu údaji o frekvenci odpovídá pouze jeden rank),
$f$ ... frekvence slova,
$k$ ... konstanta.

[89] Viz např. *Relative Frequency as a Determinant of Phonetic Change*. 1929.

[90] Více viz *The Psychology of Language. An Introduction to Dynamic Philology*. Boston 1935; *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. 1949.

**Druhý Zipfův zákon**:

$$a \cdot b^2 = k$$

(počet slov o jisté frekvenci krát frekvence na druhou je konstantní), kde
$a$ ... počet slov,
$b$ ... frekvence,
$k$ ... konstanta.

**Třetí Zipfův zákon**:

$$\frac{m}{\sqrt{f}} = k$$

(slova s vysokou frekvencí mají zpravidla větší počet významů), kde
$m$ ... počet významů daného slova,
$f$ ... frekvence,
$k$ ... konstanta.

Nyní se zastavíme u jednotlivých zákonů podrobněji.

### 2.6.1 První Zipfův zákon

Říká nám, že součin frekvence slova a jeho ranku je konstantní, tedy

$$r \cdot f = k.$$

Již v roce 1912 se ve 3. vydání svých *Gammes sténographiques*[91] pokusil francouzský stenograf J. B. Estoup formulovat zákonitost při uspořádání slov podle klesající frekvence. Zjistil, že v každém delším kontextu je součin frekvence slova $f$ a jeho ranku $r$ přibližně stejný, tedy

$$f \cdot r = k.$$

Zipf došel k podobnému závěru nezávisle na Estoupovi, proto se někdy v literatuře setkáváme s označením *zákon Estoupův-Zipfův*, častější je ale označení „*první Zipfův zákon*“.

Ve většině dostupné literatury se uvádí jako datum uveřejnění tohoto výsledku G. K. Zipfem rok 1949, W. Plath ale ve svém článku [49] uvádí, že ve skutečnosti se jednalo již o rok 1929, když ve své dizertační práci Zipf uvedl vzorec

$$X \cdot Y = n$$

(později $r \cdot f = k$) a přiložil týž graf.

Zipf vychází z hypotézy, že v jazyce existují tendence udržet rovnováhu mezi frekvencí slova a počtem slov, které tuto frekvenci mají. Jde o dvě protichůdné síly. *Síla sjednocující* (*force of unification*) působí na to, aby slova měla co nejvyšší frekvenci a jejich počet byl co nejmenší – v ideálním případě snížit počet slov na jedno s frekvencí 100 %. Proti ní působí *síla rozlišující* (*force of diversification*), jejímž cílem je, aby v jazyce bylo co nejvíce slov různých s poměrně

[91] 4. vyd. 1916, 5. vyd. 1917.

nízkou frekvencí (ideálně 1). První z nich je způsobena tzv. *ekonomií mluvčího* (a *principem minimálního úsilí* při formulaci výpovědi, který pojmenoval analogicky podle principu nejmenší akce ve fyzice), druhá pak je způsobena tzv. *ekonomií posluchače* (*minimální úsilí* při slyšení a porozumění). Tato interpretace byla ale jen zřídka přijímána jako teorie vysvětlující rozložení vztahu pořadí a četnosti.

Podle Zipfa je tedy součin frekvence slova a jeho ranku konstantní (mezi frekvencí slova a jeho rankem je nepřímo úměrná závislost), přičemž tato konstanta závisí na délce textu. Zipf jej vyjadřuje rovnicí:

$$r \cdot f = k,$$

kde
$r$ ... rank slova (pořadí slova při uspořádání slov podle klesající frekvence bez zřetele k počtu slov s touž frekvencí),
$f$ ... frekvence slova,
$k$ ... konstanta.

Zipf své teorie ověřoval na různých textech, například na románu Jamese Joyce *Odysseus* (*Ulysses*), který je tvořen 260 430 slovy, z nichž je 29 899 slov různých (průměrná slovní zásoba dospělého člověka se odhaduje na 3 000 až 10 000 slov). Při studiu tohoto textu Zipf zjistil, že součin ranků s příslušnými frekvencemi jsou přibližně stejná čísla: 26 530 (10. slovo v pořadí má frekvenci 2 653), 27 800 (50. slovo v pořadí má frekvenci 566), 26 500 (100. slovo v pořadí má frekvenci 265), 26 000 (1000. slovo má frekvenci 26) apod., tedy čísla blížící se konstantě. Zipf rovněž tvrdí, že se zákon osvědčil i při ověřování na textu publicistickém[92], při aplikaci na starou angličtinu i na texty cizojazyčné (čínština, gótština aj.).

Pravidelnost tohoto vztahu, ověřená Zipfem na textech řady jazyků, upoutala pozornost spousty badatelů, neboť se zdálo, že by mohla poskytovat klíč k nějakému obecnému principu jazykového chování. Zákon byl ale také podroben značné kritice, největší pozornost mu věnoval (zejména se zřetelem k jeho odvození) francouzský matematik B. Mandelbrot. Ukázal, „*že Zipfův vzorec sice udává „obecný spád“ křivek, ale velmi špatně zobrazuje podrobnosti*“[93]. Podrobný rozbor vztahu mezi rankem a frekvencí slova se podle něj řídí tzv. *harmonickým* a tzv. *kanonickým* zákonem.

**Harmonický zákon**:

$$p_r = \frac{P}{r},$$

kde
$r$ ... rank (pořadí),

[92] Na materiálu R. C. Eldridge, který analyzoval denní americký tisk o rozsahu 43 989 slov, z nichž bylo 6 002 slov různých.

[93] *Structure formelle des textes et communication.* Word 10, 1954, s. 1–27; český překlad *Komunikace a formální struktura textů.* In: Teorie informace a jazykověda, s. 130–150.

$p_r$ ... příslušná četnost výskytu,
$P$ ... konstanta pro každý text.
Jde v podstatě o tzv. první Zipfův zákon, kdy četnost slova je nepřímo úměrná jeho ranku.

**Kanonický zákon**:

$$p_r = P(r + \rho)^{-\beta}$$

nebo

$$\log p_r = \log P - \beta \log(r + \rho),$$

kde
$\beta$ ... konstanta pro daný text,
$\rho$ ... konstanta korigující frekvenci slov s nízkým pořadím,
$P$ ... konstanta (charakteristika rozsahu výběru),
$r$ ... rank (nabývá pro každou jednotku jiné hodnoty a je vždy nižší než počet všech různých slov, jejichž celkový počet se značí $R$),
$p_r$ ... příslušná četnost výskytu.

Harmonický zákon je speciálním případem zákona kanonického (pro $\beta = 1$ a $\rho = 0$). Mandelbrot se pokusil také kanonický zákon vysvětlit pomocí předpokladů vzatých částečně z teorie informace.

Vysvětlit rozložení $r$ a $f$, které by se opíralo o jiné hypotézy, než jaké definoval Zipf, se v pozdější době pokoušela řada autorů. Významné jsou zejména práce Mandelbrotovy, v nichž předkládá své výsledky v rámci tzv. *makrolingvistiky.* Tou rozumí nové odvětví lingvistiky, jež se má zabývat studiem jazykových jevů velkého rozsahu pomocí statistických metod. Vztah *makrolingvistika* versus *mikrolingvistika* (gramatika) je podle Mandelbrota podobný vztahu termodynamika versus mechanika jednotlivých molekul plynu: ačkoliv je popis na makroskopické úrovni slučitelný s mikroskopickým chováním popisovaným gramatikou, detailních rysů na nižší úrovni si nevšímá. I když je pouze neúplným popisem chování, může bez obtíží vést k početním výsledkům, které by prakticky nebylo možno obdržet z úvah o jednotlivých gramatických jevech. „*Mandelbrot naznačuje, že makrolingvistika může poskytnout podobný cenný nástroj pro popis hrubých rysů textového materiálu velikého rozsahu, jehož úplné, podrobné gramatické zpracování by bylo nezvládnutelně obsáhlé a složité.*“ ([49], s. 27) Navrhuje úpravu Zipfova vzorce tak, aby věrněji odpovídal pozorovaným faktům. Zavádí dva nové parametry $\rho$ a $B$ a s jejich pomocí získává tzv. *kanonický zákon*:

$$p_r = P(r + \rho)^{-B}$$

(je zde uveden v té podobě, v níž se objevuje v dřívějších Mandelbrotových článcích z let 1954 a 1955), kde $r$ značí pořadí, $p_r$ je relativní četnost slova pořadí $r$, $P, \rho, B$ jsou konstanty pro daný text ($\rho$ udává korekci pro slova nízkého pořadí).

Mandelbrotovi se podařilo odvodit kanonický zákon matematicky ze dvou různých teoretických modelů generování textů. Podle prvního jednoduššího modelu se slova textu vytvářejí písmeno za písmenem pomocí Markovova procesu

s konečným počtem stavů, v němž má každý symbol (i mezera) jistou stanovenou pravděpodobnost výskytu. Generování textu tímto způsobem vede k rozložení četností slov, které odpovídá kanonickému zákonu s $B$ větším než 1. Druhý Mandelbrotův model byl vypracován analogicky k poměrům v termodynamice; matematicky určuje „*nejpravděpodobnější stav*“ textu, který je podroben dvěma omezením: 1) všechna slova musí být během dekódovacího procesu oddělena mezerou, 2) musí být stanoven náklad optimálního dekódování (tj. systému dekódování, v němž nejkratší řada operací je přiřčena nejčastějšímu slovu atd.). Toto pojetí má za následek maximalizaci *entropie* (v duchu Shannonovy teorie informace) charakterizující rozložení pravděpodobností slov a opět vede k tomu, že rozložení četností slov je v souhlasu s kanonickým zákonem, tentokrát bez omezení pro hodnoty $B$. Mandelbrot dává zřejmě druhému modelu přednost a na jeho základě vyvozuje některé závěry – např. ten, že slova jsou základní jednotky textů, dále že teorie informace má pro lingvistiku zásadní význam.

Z dalších modelů zmiňme model Simonův (1955)[94], který považuje vytváření textu za stochastický „*proces generování*“ a na základě tohoto předpokladu odvozuje distribuční funkci, která určuje vztah mezi počtem slovních typů o dané četnosti a jejich četností výskytu. Mandelbrot[95] vytýká Simonovi to, že odvozování se děje vlastně kruhem. Belevitch[96] a Somers[97] zase vyslovují názor, že k odvození prvního Zipfova zákona stačí pouze předpoklad, že logaritmy relativních četností slov mají normální rozložení. Aproximací useknutého normálního rozložení pomocí prvního nebo prvního a druhého členu Taylorovy řady odvozuje Belevitch postupně *Zipfův zákon* a *Mandelbrotův kanonický zákon*.

Mnoho úsilí věnoval Zipfovu zákonu i G. Herdan[98], který ale odmítal mluvit o zákonu, protože už P. Guiraud[99] uvedl, že se jedná o empirickou formuli. Touto problematikou se zabýval i P. Novák[100], který tvrdí, že pravdivost takových vzorců je třeba dále prověřovat. Jedině pokud je možno takovéto vztahy odvodit, tak se zapojí do soustavy dané vědní oblasti a nezůstávají pouhými empirickými pravidly.

Prvním Zipfovým zákonem se u nás podrobně zabývala i M. Těšitelová, která se jej pokoušela ověřit pro češtinu. Ukázalo se, že příslušné vztahy platí dobře ve střední části frekvenčního slovníku, ale ne u slov s frekvencí vysokou a nízkou. K ověřování použila těchto textů:

1. K. Čapek: *Život a dílo skladatele Foltýna* (21 963 slov)

2. M. Pujmanová: *Předtucha* (24 357 slov)

[94] Simon, H. A.: *On a Class of Skew Distribution Functions*. Biometrika, sv. 42, s. 425–440.
[95] Mandelbrot, B.: *A Note on a Class of Skew Distribution Functions: Analysis and Critique of a Paper by H. A. Simon*. Information and Control, sv. 2, 1959, s. 90–99.
[96] Belevitch, V.: *On the Statistical Laws of Linguistic Distributions*. Annales de la Société Scientifique de Bruxelles, sv. 7B, 1959, s. 310–326.
[97] Somers, H. H.: *Analyse Mathematique du Langage*. Louvain 1959.
[98] Herdan, G.: *Type-Token Mathematics*. The Hague 1960.
[99] Guiraud, P.: *Problèmes et méthodes de la statistique linguistique*. Paris 1960.
[100] Novák, P.: *Význam kvantitativních metod pro lingvistiku*. In: [58].

3. I. Olbracht: *Bratr Žak* (29 803 slov)

4. *Slovník šestiletých dětí* – Praha (51 981 slov)

5. *Slovník šestiletých dětí* – Zlín (109 061 slov).

Podle šetření M. Těšitelové „*nabývá poměr mezi rankem a frekvencí slova konstantnosti v tom úseku, kde se rank slova kryje s jeho pořadím, tj. kdy za sebou následují slova s různou frekvencí a kdy se rozdíl v klesající frekvenci blíží 1. S rostoucím počtem slov s touž frekvencí, tj. se zvětšující se disproporcí mezi rankem a pořadím slova, podmíněném klesající frekvencí, hodnota „konstanty" výrazně klesá a pozbývá své platnosti zhruba u pásma slov s nejnižší frekvencí, tj. 10–1. Vzhledem k podmínce, aby se rank slova kryl s pořadím, posunuje se platnost vztahu mezi rankem a frekvencí slova mezi slova s relativně vysokou frekvencí. V jistých frekvenčních intervalech, tj. při stejném ranku, především v pásmu slov se střední frekvencí, ale i v pásmu slov s frekvencí nižší a nejnižší, vytvářejí se frekvenční roviny R, vymezené stejným rankem a stejnou frekvencí. Ty se spojují v celé frekvenční bloky B, které se vyznačují relativní stabilností ranku a frekvence, srov.*

$$B = \sum_{r_i}^{r_n} R_1 + R_2 + \ldots + R_n,$$

*kde $r_i$ je rank, od něhož začíná první frekvenční rovina, $r_n$ je rank odpovídající poslední frekvenční rovině v bloku $R_n$, $R_1$ je rovina daná rankem $r_1$ a frekvencí $f_1$* ([66], str. 46n.).“

## 2.6.2 Druhý Zipfův zákon

Vyjadřuje, že počet slov o jisté frekvenci krát frekvence na druhou je konstantní, a zapisujeme jej

$$a \cdot b^2 = k.$$

Poprvé byl uveřejněn v *The Psycho-Biology of Language* (Boston 1935). Zipf tomuto svému zákonu nepřipisuje obecnou platnost, omezuje jej pouze na slova s nepříliš velkou frekvencí, která představují značnou část slovníku. Slova s vysokou frekvencí z tohoto zákona vylučuje. Rovněž tvrdí, že tento zákon platí pro všechny jazyky.

Na českém materiálu (stejně jako u *prvního Zipfova zákona*) se pokusila M. Těšitelová ověřit jeho platnost. Ukázalo se, že tento zákon platí dobře rovněž u slov ve středním frekvenčním pásmu (a nevztahuje se ale na slova s vysokou frekvencí, ani na slova s frekvencí relativně malou). Druhý Zipfův zákon vyjadřuje v podstatě známou skutečnost, že počet slov o vysoké frekvenci je vždy malý, zatímco čím je frekvence nižší, tím větší počet slov má tuto frekvenci. B. Trnka[101] vytýká této formuli, že nepřihlíží k rozsahu textu, to znamená, že

[101] Trnka, B.: *G. K. Zipf: The Psycho-Biology of Language. G. K. Zipf: Human Behavior and the Principle of Least Effort.* ČMF 33, 1950, s. 3–4.

nemůže platit pro text libovolné délky, což ostatně M. Těšitelová na základě svých výzkumů rovněž potvrdila.

Zipf má zcela jistě pravdu v tom, že v distribuci slov existuje značná pravidelnost, která svědčí o tendencích v jazyce udržet rovnováhu mezi frekvencí slov a počtem slov různých, ovšem nelze to vyjádřit tak jednoduchou formulí.

### 2.6.3 Třetí Zipfův zákon

Říká nám, že slova s vysokou frekvencí mají zpravidla větší počet významů. Lze jej zapsat jako

$$\frac{m}{\sqrt{f}} = k.$$

Poprvé tento výsledek G. K. Zipf uvedl ve své práci *Human Behavior and the Principle of Least Effort*[102]. Matematicky se pokouší vyjádřit vztah mezi frekvencí slova a počtem jeho významů. Zipf se zde vlastně pokusil postihnout pomocí statistiky stránku sémantickou, která dosud stála stranou statistické analýzy.

I aplikace tohoto zákona na český materiál[103] ukázala, že počet významů daného slova není závislý na frekvenci slova. Např. i substantiva s frekvencí nižší a nejnižší mohou mít v podstatě týž počet významů jako substantiva s frekvencí vyšší a nejvyšší. Totéž platí u sloves. Naopak u slov formálních (např. u předložek) se ukazuje, že s klesající frekvencí ubývá předložek s relativně velkým počtem významů. Lze tedy konstatovat, že zákon platí rámcově jen pro slova formální, kterým i Zipf věnoval zvláštní pozornost.

### 2.6.4 Ostatní výsledky

Formálních slov se týká v podstatě další Zipfova empirická formule, ve které tvrdil, že frekvence slov je v nepřímém poměru k jejich délce, to znamená, že čím je slovo delší, tím má nižší frekvenci a naopak[104]. Je pravda, že slova nejvíce frekventovaná jsou poměrně krátká (srov. spojky, předložky, členy, zájmena aj.) a že je v jazyce tendence tato slova častěji opakovat. Viz například tab. 2.3 (Pořadí prvních 10 nejčastějších slov podle FSČ). Souvisí to s funkcí těchto slov, kterou lze nazvat jako funkci spojovací. Mezi délkou slova a jeho frekvencí nemůže ale existovat bezvýhradný vztah příčiny a následku, neboť v určitých komunikačních situacích volíme slova plnovýznamová, a to tak, abychom se jasně dorozuměli o určité věci.

Další myšlenka, kterou se zabýval G. K. Zipf, se týká výskytu slov v textech. Slova se v textu vyskytují v určitých *intervalech*. Jelikož se slova můžou opakovat se značnou volností, délku takového intervalu lze určit pouze přibližně. Nověji se touto problematikou zabýval i G. Herdan, který tvrdil, že to, jak často bude které slovo užito v kombinaci s jiným, je určeno kontextem. Přesto zůstaly tyto otázky jen naznačeny.

[102] Cambridge 1949.
[103] Těšitelová, M.: *On the Role of Nouns in Lexical Statistics*. PSML 2, 1967, s. 121–139.
[104] Zipf, G. K.: *Human Behavior and the Principle of Least Effort*. Cambridge 1949, s. 66n.