

Zipfov zákon v náučnom texte

Marianna Kraviarová, Filozofická fakulta PU, kraviaro@unipo.sk

Július Zimmermann, Filozofická fakulta PU, zimmer@unipo.sk

Kľúčové slová: Zipfov zákon, frekvencia slov, rank, frekvenčný slovník.

Keywords: Zipf's law, word frequency, rank, frequency dictionary.

Počiatky uplatňovania matematických metód vo všeobecnej jazykovede siahajú do prelomu 19. a 20. storočia. Išlo o kvantitatívnu lingvistiku, ktorou sa vyšetrovali frekvencie výskytu rôznych jazykových entít a javov (slov, písmen, foném, slabík, ale aj prízvukov, páuz, sonantických jadier a podobne). Je zrejmé, že porovnávacia synchronná aj asynchronná lingvistika sa dostala do popredia, pretože umožnila budovať všeobecnolingvistický obraz o svete jazykov a v neposlednom rade predstavovala rozsiahle pole pre výskumné plány. Na základe tohto priaznivého vývoja vzniká v roku 1897 prvý frekvenčný slovník, ktorého autorom je F. W. Kädling (*Häufigkeitwörterbuch der deutschen Sprache*). Slovník bol vydaný v Steglitze a obsahuje 10 910 777 slov rôznych textov, skúma frekvencie slov, slabík, slovných koreňov, ale aj prípon a predpôn. Tento frekvenčný slovník vznikol pre potreby stenografie, pre svoju rozsiahlosť a rôznorodosť jazykových jednotiek, ktoré skúmal, pomohol ku vzniku frekvenčných slovníkov pre rôzne jazyky. Dôvodom vzniku takýchto diel bola aj pomoc pri výučbe cudzích jazykov, neskôr aj rozbor štruktúry príslušného jazyka. Veľký význam mal aj výskum amerického lingvistu nemeckého pôvodu Georga Kingsley Zipfa, ktorý žil v rokoch 1902 – 1950 a v období 20. a 30. rokov 20. storočia skúmal relatívnu frekvenciu hlások. Svojim výskumom prišiel k niekoľkým veľmi významným záverom (Černý, 1996, s. 253):

- hlásky sa vyskytujú v rôznych textoch daného jazyka s rovnakou frekvenciou,
- neznelé hlásky majú výskyt asi 2x častejší ako znelé,
- čím je hláska z hľadiska artikulácie náročnejšia, tým je jej frekvencia nižšia.

Zipf matematicky formuloval uvedené frekvenčné javy v jazyku, vo všeobecnosti sú známe ako Zipfove zákony. Prvý Zipfov zákon formalizuje vzťah medzi poradovým číslom slova vo frekvenčnom slovníku (rankom slova) a jeho frekvenciou v analyzovanom texte. Podľa tohto zákona je medzi týmito dvomi veličinami nepriama úmera – čím je rank slova nižšie číslo, tým je jeho frekvencia vyššia a naopak. Platí vzťah:

$$r \cdot f = k$$

r – rank slova v slovníku s klesajúcou frekvenciou

f – frekvencia slova v texte

k – konštanta v ohraničenom intervale

Na experimentovanie so Zipfovým zákonom sa používajú jednoúčelové frekvenčné zoznamy, alebo frekvenčné slovníky. Jedným zo základných problémov pri štatistických výskumoch je výber správnej vzorky – reprezentatívneho súboru, ktorý by mal obsahovať čo najviac druhov a typov z jazykového prejavu. V roku 1969 vyšiel frekvenčný slovník slovenského jazyka pod názvom *Frekvencia slov v slovenčine* od autora Jozefa Mistríka v rozsahu 1 000 000 slov, zozbieraných v rokoch 1922 – 1966. Ide o slovník stredného typu, ktorý pozostáva z 5 štylistických skupín:

1. dialógy
2. umelecká próza

3. poézia
4. žurnalistika
5. náučná próza

Pri overovaní platnosti prvého Zipfovho zákona sme porovnali zistené frekvencie slov vybraného diela s týmto frekvenčným slovníkom.

Frekvenčný zoznam sme vytvorili z monografie Juraja Rusnáka – *Textúry elektronických médií (Vývoj a súčasný stav)*. Z tejto publikácie sme vybrali text 3. – 6. kapitoly, ktorý obsahoval 2751 rozličných slov v úhrnnom počte 12 979 slov. Vynechali sme cudzie slová v poznámkach pod čiarou, číslice, skratky; slová sme modifikovali do základného tvaru (nominatív singuláru), slovesá do neurčitku a predložky *v* – *vo*, *k* – *ku*, *s* – *so*, *z* – *zo* sme spojili do jedného významu. Takýmto spôsobom sme vytvorili frekvenčný zoznam s klesajúcou frekvenciou. Keďže skúmaný text je monotematický, orientovaný na mediálny odbor, frekvenčný zoznam nie je totožný s frekvenčným slovníkom slovenčiny.

V tabuľke 1 uvádzame prvých 50 najfrekventovanejších slov v analyzovanom texte. Jednotlivé stĺpce tabuľky obsahujú poradové číslo slova vo frekvenčnom zozname, rank daného slova a jeho absolútnu frekvenciu.

p. č.	rank	slovo	frekvencia	p. č.	rank	slovo	frekvencia
1	1	v (o)	563	26	23	do	63
2	2	a	449	27	24	komunikačný	62
3	3	byť	307	28	24	pre	62
4	4	sa	210	29	25	podobne	58
5	5	ktorý	175	30	25	relácia	58
6	6	na	172	31	26	on	56
7	7	program	145	32	27	prostredie	55
8	8	médium	140	33	28	tzv	53
9	9	rozhlasový	124	34	29	alebo	50
10	10	televízny	122	35	30	možno	48
11	11	mediálny	117	36	30	tento	48
12	12	aj	116	37	31	resp	46
13	13	pri	100	38	32	k (u)	45
14	14	elektronický	96	39	33	dielo	43
15	14	z (o)	96	40	34	znak	42
16	15	ako	87	41	35	mať	41
17	15	vysielanie	87	42	36	časť	38
18	16	text	79	43	36	rok	38
19	17	o	75	44	37	televízia	37
20	17	s (o)	75	45	38	formát	36
21	18	programový	72	46	38	stanica	36
22	19	informácia	71	47	38	štruktúra	36
23	20	komunikácia	68	48	38	typ	36
24	21	napríklad	67	49	39	ale	35
25	22	hudobný	64	50	39	jednotlivý	35

Tabuľka 1 – frekvenčný zoznam slov analyzovaného textu.

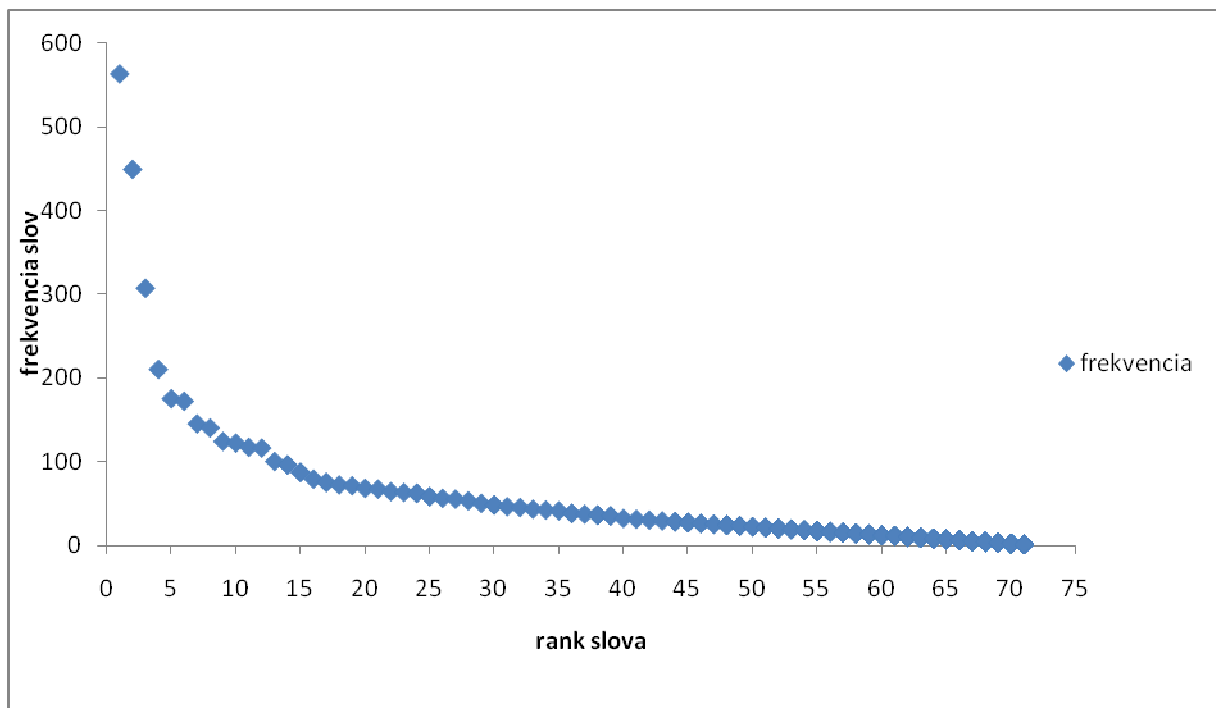
Pre porovnanie s frekvenčným slovníkom slovenčiny (Mistrík, 1969) uvádzame v tabuľke 2 zoznam 20 najfrekventovanejších slov:

p. č.	rank	slovo	p. č.	rank	slovo
1	1	a	11	11	do
2	2	byť	12	12	komunikačný
3	3	v	13	13	pre
4	4	na	14	14	podobne
5	5	sa	15	15	relácia
6	6	ten	16	16	on
7	7	on	17	17	prostredie
8	8	že	18	18	tzv
9	9	z	19	19	alebo
10	10	ako	20	20	možno

Tabuľka 2 – najfrekvencovanejšie slová z frekvenčného slovníka slovenčiny.

Porovnaním oboch tabuliek sa dá zistiť, že na popredných miestach podľa frekvenčného výskytu sa nachádzajú rovnaké slová – ide o neohybné slová (predložky, spojky), a z ohybných slov zámená a pomocné slovesá. Až na ďalších miestach stoja plnovýznamové slová. Pretože analyzovaný text nemá veľký rozsah a je odborným textom zameraným na určitú oblasť, odrazilo sa to aj v najfrekvencovanejších slovách príslušného textu.

Grafické znázornenie klesajúcej frekvencie slov s ich stúpajúcim rankom môžeme sledovať na grafe 1 – krivka má exponenciálny tvar.



Graf 1 – závislosť frekvencie slov na ranku slov.

Podľa prvého Zipfovho zákona súčin frekvencie slova a jeho ranku v danom frekvenčnom zozname je približne konštantný. Tento zákon neplatí pre slová z okrajových

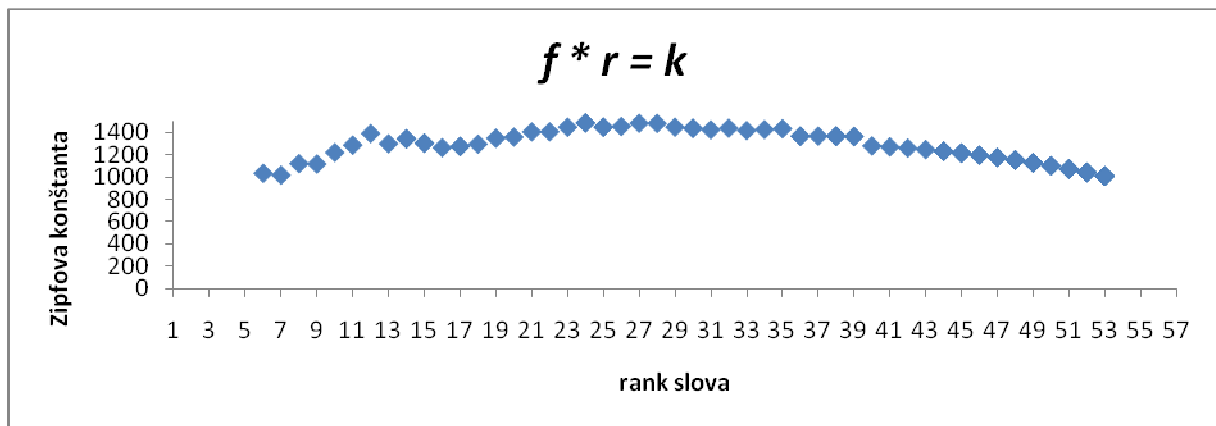
časti frekvenčného slovníka. Ak vyčíslime konštanty k pre slová s rankom r 6 – 53, zistíme, že závislosť frekvencie slova na jeho ranku je veličina s malým rozptylom. Hodnoty frekvencie slov a konštant k uvádzame v tabuľke 3.

p. č.	rank	slovo	frekvencia	$k = f * r$	p. č.	rank	slovo	frekvencia	$k = f * r$
6	6	na	563	1032	54	43	taký	29	1247
7	7	program	449	1015	55	44	od	28	1232
8	8	médium	307	1120	56	44	prax	28	1232
9	9	rozhlasový	210	1116	57	44	svoj	28	1232
10	10	televízny	175	1220	58	45	čas	27	1215
11	11	mediálny	172	1287	59	45	film	27	1215
12	12	aj	145	1392	60	45	ich	27	1215
13	13	pri	140	1300	61	45	používať	27	1215
14	14	elektronický	124	1344	62	46	vytváranie	26	1196
15	14	z (o)	122	1344	63	46	za	26	1196
16	15	ako	117	1305	64	47	system	25	1175
17	15	vysielanie	116	1305	65	47	zvyčajne	25	1175
18	16	text	100	1264	66	48	často	24	1152
19	17	o	96	1275	67	48	hudba	24	1152
20	17	s (o)	96	1275	68	48	kód	24	1152
21	18	programový	87	1296	69	48	podľa	24	1152
22	19	informácia	87	1349	70	48	tak	24	1152
23	20	komunikácia	79	1360	71	48	že	24	1152
24	21	napríklad	75	1407	72	49	prenos	23	1127
25	22	hudobný	75	1408	73	49	proces	23	1127
26	23	do	63	1449	74	49	rozhlas	23	1127
27	24	komunikačný	62	1488	75	49	udalosť	23	1127
28	24	pre	62	1488	76	50	druh	22	1100
29	25	podobne	58	1450	77	50	teória	22	1100
30	25	relácia	58	1450	78	50	už	22	1100
31	26	on	56	1456	79	51	krajina	21	1071
32	27	prostredie	55	1485	80	51	najčastejší	21	1071
33	28	tzv	53	1484	81	51	programovanie	21	1071
34	29	alebo	50	1450	82	51	tvorba	21	1071
35	30	možno	48	1440	83	51	vysielací	21	1071
36	30	tento	48	1440	84	51	znakový	21	1071
37	31	resp	46	1426	85	52	audiovizuálny	20	1040
38	32	k (u)	45	1440	86	52	či	20	1040
39	33	dielo	43	1419	87	52	označenie	20	1040
40	34	znak	42	1428	88	52	rituál	20	1040
41	35	mať	41	1435	89	52	správa	20	1040
42	36	časť	38	1368	90	52	šírený	20	1040
43	36	rok	38	1368	91	52	verejnoprávny	20	1040

44	37	televízia	37	1369	92	52	vznikať	20	1040
45	38	formát	36	1368	93	53	až	19	1007
46	38	stanica	36	1368	94	53	dôležitý	19	1007
47	38	štruktúra	36	1368	95	53	fungovanie	19	1007
48	38	typ	36	1368	96	53	jeden	19	1007
49	39	ale	35	1365	97	53	keď	19	1007
50	39	jednotlivý	35	1365	98	53	komerčný	19	1007
51	40	medzi	32	1280	99	53	spoločnosť	19	1007
52	41	publikum	31	1271	100	53	všetok	19	1007
53	42	iný	30	1260	101	53	zákon	19	1007

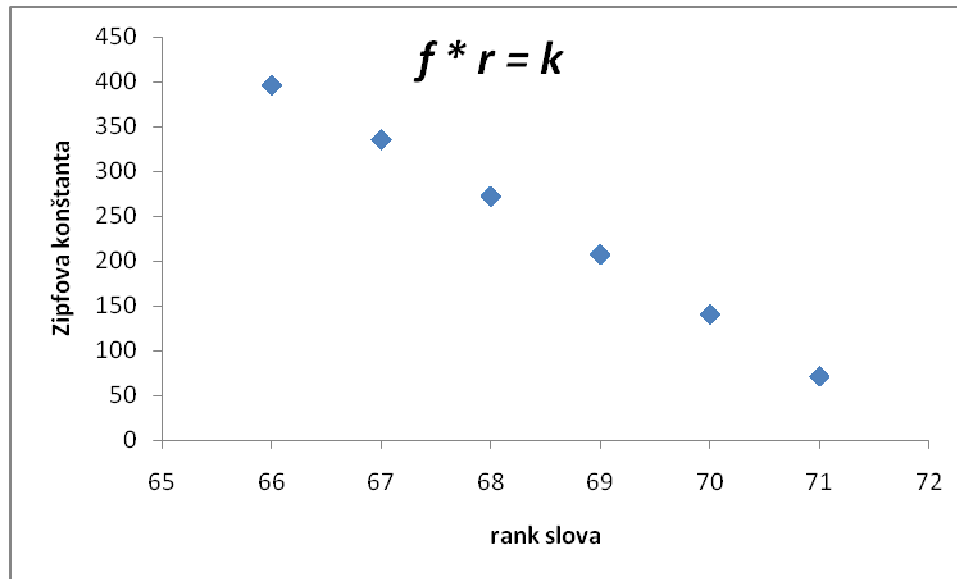
Tabuľka 3 – hodnoty frekvencie slova a ranku slova.

Závislosť medzi frekvenciou slova a rankom slova je vyjadrená graficky v grafe 2.



Graf 2 – závislosť medzi frekvenciou slova a rankom slova.

Vo frekvenčnom zozname rozlišujeme 3 pásma slov. Prvé pásmo – najvyššie – obsahuje najfrekventovanejšie slová, v analyzovanom texte tvoria cca 14% textu. Druhé pásmo – stredné – obsahuje slová strednej frekvencie, v danom texte ide o slová s rankom 6 – 53, sú to slová s konštantou $k = (1007;1488)$. Tretie pásmo – najnižšie – obsahuje slová s nízkou až najnižšou frekvenciou. V analyzovanom texte až 1439 slov z 2751 má frekvenciu 1 a 454 slov má frekvenciu 2. Slová nachádzajúce sa v treťom pásme hovoria aj o bohatstve slovníka. Závislosť medzi rankom a frekvenciou slova v treťom pásme s výskytom 1 až 6-krát je znázornená v grafe 3. Táto závislosť má klesajúcu tendenciu.



Graf 3 – závislosť medzi frekvenciou slova a rankom slova.

Záver: Experimentálne sme overili platnosť prvého Zipfovho zákona: frekvencia slova v texte je inverzne proporcionálna jeho poradiu vo frekvenčnom slovníku. Vzťah o konštantnom súčine frekvencie slova a jeho ranku platí pre stredné pásmo frekvenčného zoznamu. V okrajových pásmach uvedeného zoznamu má táto závislosť silne klesajúcu tendenciu.

Literatúra

ČERNÝ, J.: *Dějiny lingvistiky*. Olomouc, Votobia 1996. 513 s.

JELÍNEK, J., BEČKA, J. V., TĚŠITELOVÁ, M.: *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha, SPN 1961. 588s.

MISTRÍK, J.: *Frekvencia slov v slovenčine*. Bratislava, Vydavateľstvo SAV 1969. 728 s.

RUSNÁK, J.: *Textúry elektronických médií (Vývoj a súčasný stav)*. Prešov, Filozofická fakulta Prešovskej univerzity v Prešove 2009. 295 s.

TĚŠITELOVÁ, M.: *Kvantitativní lingvistika*. Praha, SPN 1987, s. 42 – 67.

Abstract

This paper deals with the validity of the first Zipf's law; it means that between order number of a word in frequency dictionary and its frequency in a text there is reciprocal proportion. As a researched text it is 40 paged scientific monography. Presented reciprocal proportion is shown by x-y graph.

„Táto štúdia, bola vytvorená realizáciou projektu *Vybudovanie lingvokulturologického a prekladateľsko-tlmočnického centra*, na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja.“