UDK 811.163.6:681.3
Peter Grzybek
Univerza v Gradcu

POGOSTNOSTNA ANALIZA BESED IZ ELEKTRONSKEGA KORPUSA SLOVENSKIH BESEDIL

V tem prispevku gre za kvantitativno analizo besedne dolžine in pogostnosti v slovenščini ter s tem povezanimi vprašanji. Analiza sicer znatno presega okvire praktične uporabnosti, hkrati pa je toliko ponazorjevalna, da lahko programsko nakaže različne raziskovalne možnosti, ki se v slovenskem jezikoslovju in v literarni vedi doslej še niso uporabljale.

The article deals with quantitative analysis of lexical length and frequency in Slovene and with related questions. The analysis goes beyond the limits of practical use, but at the same time it is illustrative enough to programmatically point out various research options that have so far not been used in Slovene linguistics and literary criticism.

0. Uvod

Pri analizi bodo različna vprašanja nujno (že iz prostorske stiske) ostala ob strani. Tako se moramo odreči npr. *besedilni* analizi, namesto tega pa se bomo posvetili besedni ravni v obliki besednega seznama, neodvisnega od besedilne ravni, se pravi, seznama, v katerem je vsaka besedna oblika zapisana točno enkrat. Analize, ki upoštevajo (tudi) stavčno in besedilno raven, bo treba opraviti drugje.

Za gradivo bo služil pogostnostni seznam 1000 najpogostejših slovenskih besed iz elektronskega korpusa slovenskih besedil (CORTES), ki ga je sestavil Primož Jakopin. Ta korpus je v času analize (december 1999) sestavljajo 112 literarnih (pretežno proznih) besedil iz 19. in 20. stoletja. Besedila so dela 41 avtorjev, pri čemer je 98 besedil izvirno slovenskih, v 14 primerih pa gre za prevode v slovenščino.[1]

Vnaprej je treba poudariti razliko med analizo besednega seznama na eni strani in besedno analizo kot celostno sestavino besedilne analize na drugi. Zavedati se moramo namreč, da oba postopka dajeta preveč različne rezultate, kar je povezano s tem, da na besedno strukturo znotraj nekega besedila vpliva pač (tudi) vsakokratna besedilna struktura.

Omejitev na seznam besedne frekvence je načeloma mogoča in tudi ni problematična; tovrstna analiza je v določenem smislu (glede na status in kakovost raziskovanega gradiva) primerljiva z običajnimi slovarskimi analizami. Razlika je le v tem, da pogostostni seznam (tj. pogostnost pojavljanja) pogojuje ***poseben izbor***, kar je primerljivo z analizo pogostostnih slovarjev.

---

[1] Za to analizo je bilo iz seznama 1000 najpogostejših izločenih devet tujih enot, ki so v korpus prišle domnevno s prevodi, za katere pa ne moremo trditi, da ne bi popačili celotne slike: *Bouvard*, *Fogg*, *IBM*, *Maltzahn*, *Passepartout*, *Pécuchet*, *Ray*, *Timmy*, *Winston*. – V besedilu pa je ostala vrsta slovenskih lastnih imen, čeprav je o lastnih imenih znano, da je njihova srednja dolžina v poprečju daljša od srednje dolžine samostalnikov (prim. Jakopin 1996).

Prav zaradi tega dejstva pa se srečujemo z ne nepomembnim problemom, ki se ga moramo zavedati: Pri našem frekvenčnem seznamu nimamo opravka z gradivom, ki bi temeljilo zgolj na enem besedilu in pri katerem frekvenca posameznih besed ne bi bila pomembna. Ta seznam namreč temelji na večjem številu različnih besedil. Zato pri raziskovanem besedilnem korpusu in posledično tudi pri našem frekvenčnem seznamu ne gre za homogeno podatkovno gradivo, temveč za rezultat raznovrstnih besedil.

Glede na besedila oz. besedilne korpuse je Orlov (1982) taka heterogena besedila označil za »kvazibesedila« in navedel prepričljive utemeljitve (in zbral tudi empirične dokaze) za to, da v jeziku ni take besedilne enotnosti, ki bi bila dovolj enovita, da bi imela konstantne parametre – saj še celo posamezna besedila niso nujno enovita. Zato so se v zadnjem času v kvantitativnem jezikoslovju odrekli statistiki *jezika* kot sistema v korist statistike *govora*.

Altmann (1992: 287) je to problematiko razširil tudi na analize slovarjev oz. besednih seznamov in glede na to upravičeno poudaril, da je celo frekvenca posamezne besede mešanica heterogenih sekvenc. Avtor ugotavlja: »Domneva, da se približujemo jezikovni normi, če zberemo veliko podatkov, je napačna. Na pogostnost [...] vpliva toliko lokalnih dejavnikov (npr. avtor, stil, besedilna vrsta itd.), da bi bilo povsem iluzorno govoriti o pogostnosti besede v jeziku«.

V tem pogledu se zgornja formulacija našega vprašanja, se pravi, vprašanja besedne dolžine v »slovenščini«, izkaže za napačno. Na podlagi dejstva, da torej nimamo opraviti z analizo posameznih besedil oz. z analizami besedne pogostnosti, ki bi temeljili na posameznih besedilih, temveč z raziskavo seznama besedne pogostosti, ki temelji na več besedilih, smo torej pri leksikalni analizi soočeni s problemom podatkovne neenotnosti; temu problemu se tudi ne moremo izogniti.

Načeloma imamo torej pri našem seznamu iz CORTES-a v dvojnem pogledu opraviti s posebnim izborom: Na eni strani gre, kot smo že povedali, za besede z zelo visoko pogostnostjo, na drugi pa imamo opraviti s seznamom besedne pogostnosti, ki temelji na mešanem korpusu in ki je ne gre izenačevati s seznamom besedne pogostnosti, ki temelji na posameznem besedilu. Dvojno posebnost izbora tako določa dejavnik ***transbesedilne frekvence***.

Vendar pa pravkar opisana posebnost gradiva – če se je le zavedamo – ponuja celo vrsto analiz, ki jih bomo v nadaljevanju pokazali. Konkretno gre za raziskave o naslednjih vprašanjih:

1. pogostnosti besed in njeni porazdelitvi
2. razmerju med dolžino besed in zlogov (neodvisno od pogostnosti pojavljanja vsakokratnih enot v besedilih)
3. besedni dolžini (ob upoštevanju, kako pogosto se enote pojavljajo v besedilih)
4. pogostnosti besedne dolžine in njeni porazdelitvi

Začnimo našo analizo z raziskavo besedne pogostnosti.

1. Besedna pogostnost

Prav na podlagi dejstva, da naš seznam besedne pogostnosti temelji na mešanem korpusu, se zdi smiselno, da se najprej lotimo tega vprašanja – ki, mimogrede, (še) nima opraviti z vprašanji besedne dolžine in njene pogostnosti, temveč s pogostnostjo besed (oz. besednih oblik). Praviloma namreč pogostnosti, s katero se besede pojavljajo v posameznih besedilih, ni naključna (kaotična), ampak je posledica določene zakonitosti. V zvezi z našim seznamom besedne pogostnosti bi bilo treba preveriti, ali porazdelitev besedne pogostnosti v njem ustreza porazdelitvi besedne pogostnosti v posameznih besedilih. Da ne bo nesporazuma, še enkrat poudarimo, da ne vprašujemo, katere besede se pojavljajo s katero pogostnostjo, ampak kako visoka je absolutna pogostnost najpogostejše, druge najpogostejše, tretje najpogostejše itd. besedne oblike, ne glede na to, za katere konkretne besede pri tem gre.

Ne sprašujemo le, kako pogosto se pojavljajo besede, temveč tudi, ali lahko iz tega izhajajočo pogostnostno porazdelitev teoretično modeliramo, se pravi, ali lahko izpeljemo matematični izrek, na podlagi katerega bi izdelali (teoretični) porazdelitveni model, s katerim bi lahko izračunali in modelirali pogostnost. Za besedno pogostnost je primeren model tako imenovane Zipf-Mandelbrotove porazdelitve. Model temelji na G. K. Zipfovem načelu (iz 30. in 40. let), ki ga je v 50-ih letih modificiral Mandelbrot in ki pravi, da produkt iz ranga leksikalne enote v naključnem vzorcu in njene dejanske (absolutne) frekvence da konstanto. V nadaljnji izpeljavi in modifikaciji te domneve po B. Mandelbrotu se da po formuli

$$(1) \qquad P_x = \frac{(b+x)^{-a}}{F(n)} \quad \text{z: } x = 1,2,3,\ldots,n; \quad a,b > 0;\ n\,\varepsilon\,\subseteq; \quad F(n) = \sum_{i=1}^{n} (b+i)^{-a}$$

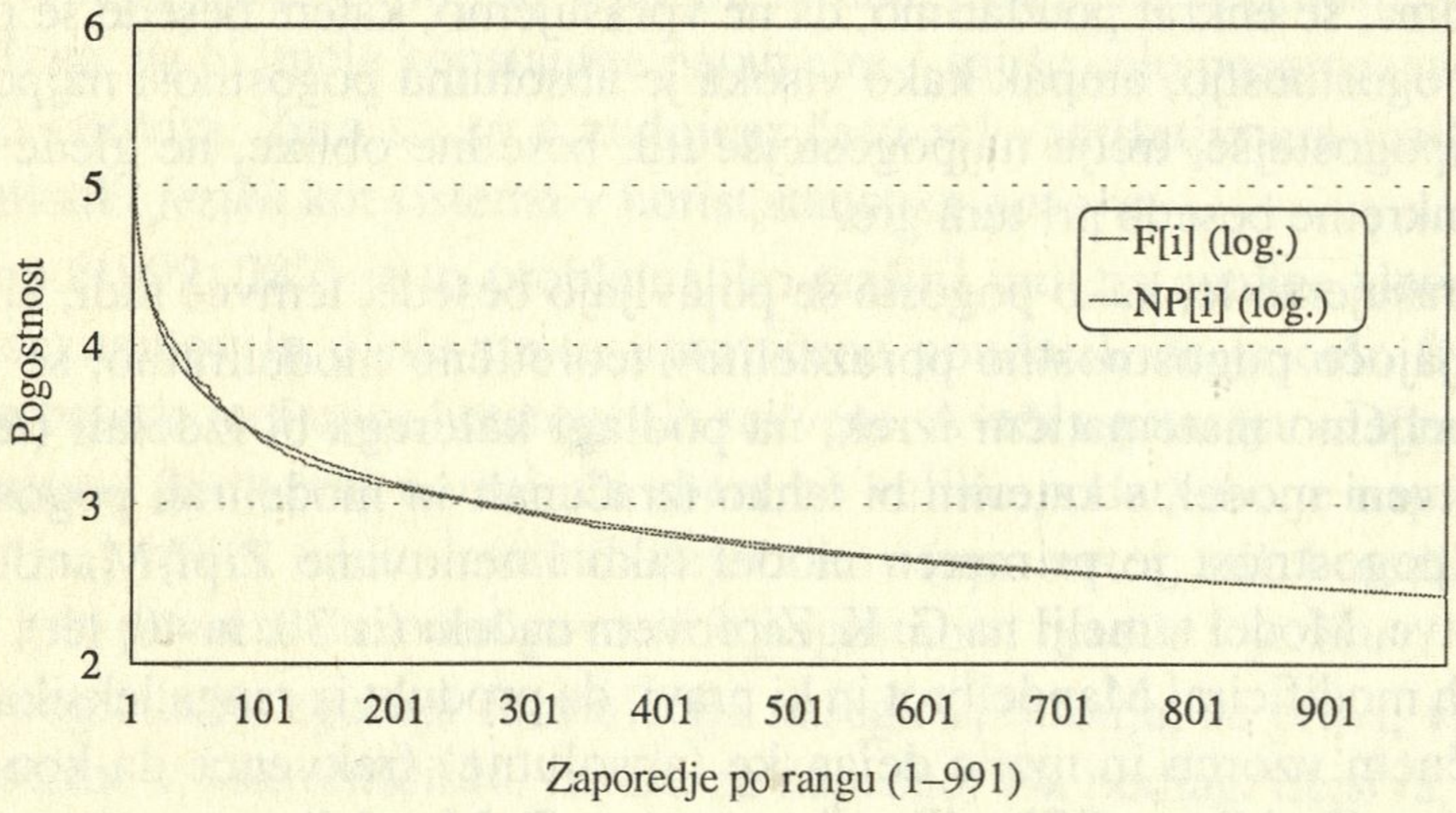
teoretično izračunati pogostnost pojavljanja vsakokratnih leksemov: – *a* in *b* sta pri tem (od vzorca do vzorca različna) parametra, *F(n)* je normirna konstanta, ki se (kot je opisano) izračuna iz obeh omenjenih parametrov in vsakokratnega ranga določene enote. V kolikšni meri se zgornja formula, po kateri dobimo teoretične (ocenjene vrednosti), ujema z dejanskimi vrednostmi, nato praviloma preverimo s tako imenovanim χ^2-testom. Ker pa je ta pri velikih vzorcih (s takimi pa imamo pri jezikovnem gradivu pogosto opravka) sorazmerno hitro pomemben, se pri vzorcih z velikim *N* namesto tega uporablja tudi kot χ^2/N izračunani kontingenčni koeficient *C*. Ta koeficient velja pri $C < 0.02$ za kazalnik dobrega, pri $C < 0.01$ kot zelo dobrega približka (Grotjahn/Altmann 1993) – v tem primeru, z drugimi besedami, velja, da je teoretični izračun primeren za to, da se empirično ugotovljene vrednosti zajamejo v splošnem modelu.

Ker je za naš seznam besedne pogostnosti znana pogostnost pojavljanja posameznih besednih oblik, lahko le-te razvrstijo vrednosti v ustrezno zaporedje po rangu. Ob padajočem razvrščanju so tako najpogostejše besede na prvih, (relativno) manj pogoste na spodnjih mestih. Glede na pogostnostni seznam po rangu, na katerem je 991 leksikalnih enot, dobimo z vstavitvijo parametrov *a* =

1.007 in $b = 0.5004$ v formulo (1) Zipf-Mandelbrotove formule (1) formulo (1a), ki pri kontingenčnem koeficientu $C = 0.0082$ dejansko pokaže izvrsten približek (ker $p < 0.01$).

$$(1a) \quad P_x = \frac{(0.5004 + x)^{-1.007}}{F(991)}$$

Zaradi prostorske omejitve se moramo odreči natančni predstavitvi rezultatov. Namesto tega so v grafu 1 prikazani izmerjeni in teoretični podatki, pri čemer so pogostnostne vrednosti zaradi preglednosti predstavljene logaritmično.



Graf 1: Ujemanje Zipf-Mandelbrotove porazdelitve s pogostnostjo po rangu

Kot je videti, se krivulji, ki izhajata iz empiričnih in teoretičnih vrednosti, skoraj pokrivata. Lahko torej domnevamo, da pogostnostna porazdelitev iz našega seznama, ki temelji na mešanem korpusu, ustreza tistemu porazdelitvenemu modelu, značilnem tudi za porazdelitev besedne pogostnosti v posameznih besedilih.

Ta rezultat bi lahko povsem upravičeno utemeljeval enako nadaljevanje tudi pri drugih analizah, kakor bi ravnali s seznamom besedne pogostnosti v posameznem besedilu. Vendar je kljub pozitivnemu izsledku treba biti previden pri njegovi interpretaciji: celo če je za naš seznam Zipf-Mandelbrotova porazdelitev odličen približek, se s teoretičnega vidika ne da izključiti, da imamo opraviti samo z umetnim rezultatom, pogojenim s slučajnim seštevkom delnih rezultatov.[2] V tem pogledu te domneve ne bomo mogli niti dokazati niti ovreči, zato se raje posvetimo dejanski analizi besedne dolžine.

2. Razmerje med dolžino besed in zlogov

V jezikoslovju na splošno in posebej v stilistiki (kot prehodu v literarno vedo) je izračunavanje povprečnih besednih dolžin (tj. srednjih vrednosti besednih

[2] Že iz tega razloga je treba v CORTES-u predlagane možnosti analize brezpogojno razširiti oz. pravzaprav omejiti, namreč na obseg posameznih besedil.

dolžin) povsem običajen postopek. Ta med drugim velja tudi za pogoj, da se srednje vrednosti dveh različnih naključnih vzorcev (npr. dveh različnih besedil enega avtorja ali dveh različnih avtorjev) medsebojno primerjajo. Tudi tu se bomo v nadaljevanju, čeprav drugače, ukvarjali s tem vprašanjem. Pred tem pa velja opozoriti na drugo, z besedno dolžino povezano vprašanje, s katerim se lahko navežemo na ustrezna dela iz drugega konteksta (Grzybek 1999), kar nam bo omogočilo boljšo umestitev rezultatov za slovenščino.

Izhodišče tega vprašanja je domneva, da dolžina besed v besedilih ni izolirana, ampak povezana z drugimi prvinami, in sicer po eni strani z dolžino stavkov, po drugi z dolžino zlogov. Ker se bomo pri tej raziskavi zadržali le na besedni ravni, ne bomo upoštevali dolžine stavkov oz. odvisnikov kot (možnega) korelacijskega dejavnika, zato se bomo posvetili razmerju med dolžino besed in zlogov.

To vprašanje se navezuje na dela iz 50-ih let, npr. na Gajićevo disertacijo (1950), kjer je izračunana povprečna dolžina zlogov vseh gesel v Junckerjevem Nemško-hrvaškem slovarju iz leta 1930. Pri tem se je jasno pokazalo, da večanje zlogov v besedah vpliva na to, da so zlogi v povprečju krajši. Bonnski romanist Paul Menzerath, pri katerem je Gajić pisal disertacijo, je to tendenco v svojem delu *Architektonik des deutschen Wortschatzes / Arhitektonika nemškega besedišča* (1954: 101) prevedel v posplošitev: »Relativno število glasov se ob rastočem številu zlogov zmanjšuje, ali z drugimi besedami, več zlogov ko ima beseda, toliko (relativno) krajša je.«

Altmann (1980), ki je to trditev kasneje povzdignil v status tako imenovanega Menzerathovega zakona, je ta zakon razširil na vse jezikovne ravnine, tako da se v posplošeni obliki glasi: »Kolikor večja oz. kompleksnejša je jezikovna tvorba, toliko manjši oz. enostavnejši so njeni sestavni deli.« Altmann je razen tega izpeljal matematično formalizacijo te težnje v obliki nelinearne regresije, ki se v najsplošnejši obliki (2) glasi:

(2) $y = ax^b\, e^{-cx}$

s posebnima primeroma (2a) in (2b):

(2a) $y = ax^b$ za $b \neq 0, c = 0$

(2b) $y = ae^{-cx}$ za $b = 0, c \neq 0$.

Te tri formule, ki jih poznamo tudi kot Menzerath-Altmannov zakon, so prikazane v tabeli 1.

Tab. 1: Menzerathov zakon v Altmannovi formalizaciji (1980)

I	$b = 0$		$y = Ae^{-cx}$
II	$b \neq 0$	$c = 0$	$y = Ax^b$
III		$c \neq 0$	$y = Ax^be^{-cx}$

Glede pokrivanja teoretičnih modelov z izmerjenimi podatki se v primerjavi s posebnima primeroma (I) in (II) običajno najboljši približek doseže z najkom-

pleksnejšo formulo (III), kar je seveda povezano s tem, da izkazuje največ parametrov (*a*, *b* in *c*). Praviloma pa si vendarle prizadevamo, da za uspešne(jše) približke štejemo tiste, pri katerih teoretični izračun zmanjšamo na čim manjše število parametrov. V tem smislu šteje formula (II) običajno za »standardni primer« Menzerath-Altmannovega zakona. S to formulo je tudi Altmann (1989: 55) prilagodil Gajićeve podatke in dejansko dobimo z njo odlične rezultate. Hkrati je treba povedati, da se uspešnost približka ocenjuje s tako imenovanim determinacijskim koeficientom R^2, mero za razmerje med izmerjenimi in teoretičnimi vrednostmi v intervalu med 0 in 1. Na tem mestu se lahko izognemo podrobnostim, ker so rezultati ponovne analize Gajićevih podatkov in ujemanja formule (II) s temi podatki obširneje opisani na drugem mestu (prim. Grzybek 1999). Če povzamemo izsledke, naj omenimo le, da je determinacijski koeficient v primeru opisanega približka le minimalno pod najvišjo vrednostjo 1, namreč pri $R^2 = .977$ – to pomeni, da je povprečna dolžina zloga v skoraj 98 % določena z dolžino besede. Ustrezna regresijska enačba (prim. 2a) se glasi $y = 3.23x^{-0.278}$, vrednosti po tej enačbi so v četrtem stolpcu tabele 2. Zanimivo pa je dejstvo, da je Grzybek (1999) z drugačno regresijsko enačbo dosegel enako dober, celo za spoznanje boljši približek, namreč s standardno enačbo (3) iz kompleksnega statističnega programa SPSS:

(3) $y = e^{(a+bx)}$

Enačbo (3) so sicer med drugim že uporabljali za opis razmerja med dolžino besed in pomenskim obsegom, vendar so do zdaj razmerja med tvorbo in sestavnimi deli raje opisovali s formulo (II) Menzerath-Altmannovega zakona.

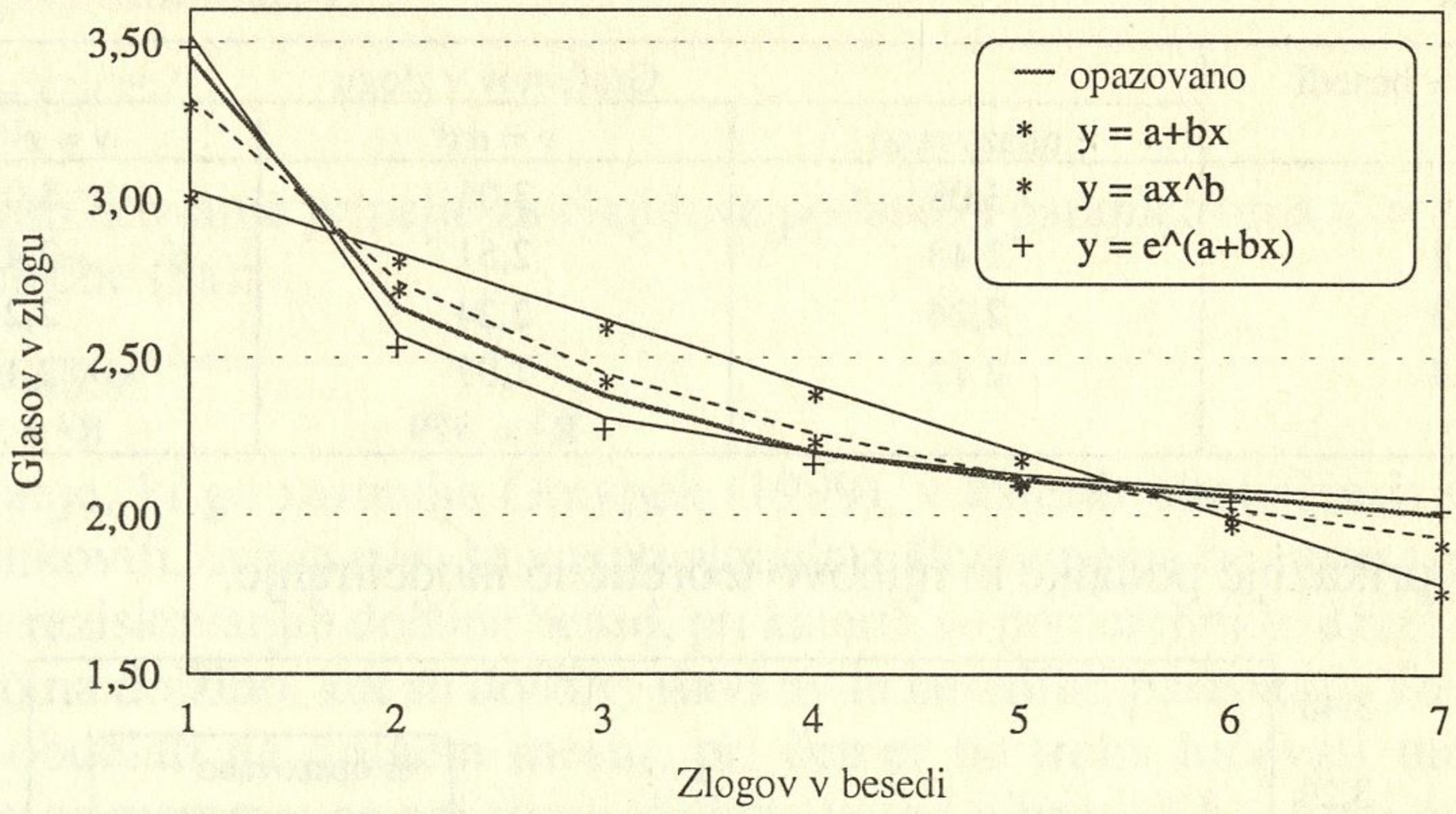
Po vstavitvi parametrov $a = 0.628$ in $b = 0.630$ dobimo enačbo (3a) z $R^2 = .985$ – ustrezne teoretične vrednosti so v petem stolpcu tabele 2.

(3a) $y = e^{(0.628+0.630x)}$

Tabela 2: Teoretični približki med dolžino besed in zlogov

Zlogov v besedi	Število besed s *f* zlogov v besedi	Glasov v zlogu (opazovano)	Glasov v zlogu (teoretično)	Glasov v zlogu (teoretično)
f	*n*		$\overline{x}$ $[y = ax^b]$	$\overline{x}$ $[y = e^{(a+bx)}]$
1	717	3,45	3,32	3,49
2	4038	2,66	2,74	2,57
3	6060	2,38	2,45	2,32
4	5066	2,20	2,26	2,20
5	1239	2,11	2,12	2,14
6	145	2,06	2,02	2,09
7	14	2,00	1,93	2,06
			$R^2 = .977$	$R^2 = .987$

Graf 2 kaže ujemanje treh izračunanih modelov.



Graf 2: Modeli približkov za dolžino besed in zlogov

Tako se je dalo v raziskovanem hrvaškem slovarskem gradivu ugotoviti razmerje med dolžino besed in zlogov, ki ga lahko približno enako dobro izračunamo z dvema regresijskima enačbama (2a, 3). Grzybek (1999) se je zavzemal, da se v nadaljnjih raziskavah odkriva, pod katerimi pogoji sta enačbi učinkoviti oz. pod katerimi pogoji se ena izmed obeh izkaže za boljšo. Na tem mestu se zdi primerno to vprašanje ponoviti in ga projicirati na naš seznam besedne pogostnosti iz CORTES-a.

Poskusimo torej odgovoriti na vprašanje, kako je z razmerjem med dolžino besed in zlogov pri slovenskem seznamu besedne pogostnosti. Kot kažejo podatki v tabeli 3, se zgoraj ugotovljena osnovna tendenca potrjuje tudi pri slovenskem gradivu: Daljša ko je (po zlogih) beseda, (relativno) krajši so (po grafemih) zlogi. Po pričakovanju pokaže za Menzerath-Altmannov zakon uporabljena enačba

(2a) $y = ax^b$

s parametroma $a = 3.05$ in $b = 0.28$ pri $R^2 = .98$ zelo dober približek. Pomembno pa je naslednje: Kot tudi že pri hrvaškem gradivu pokaže regresijska enačba

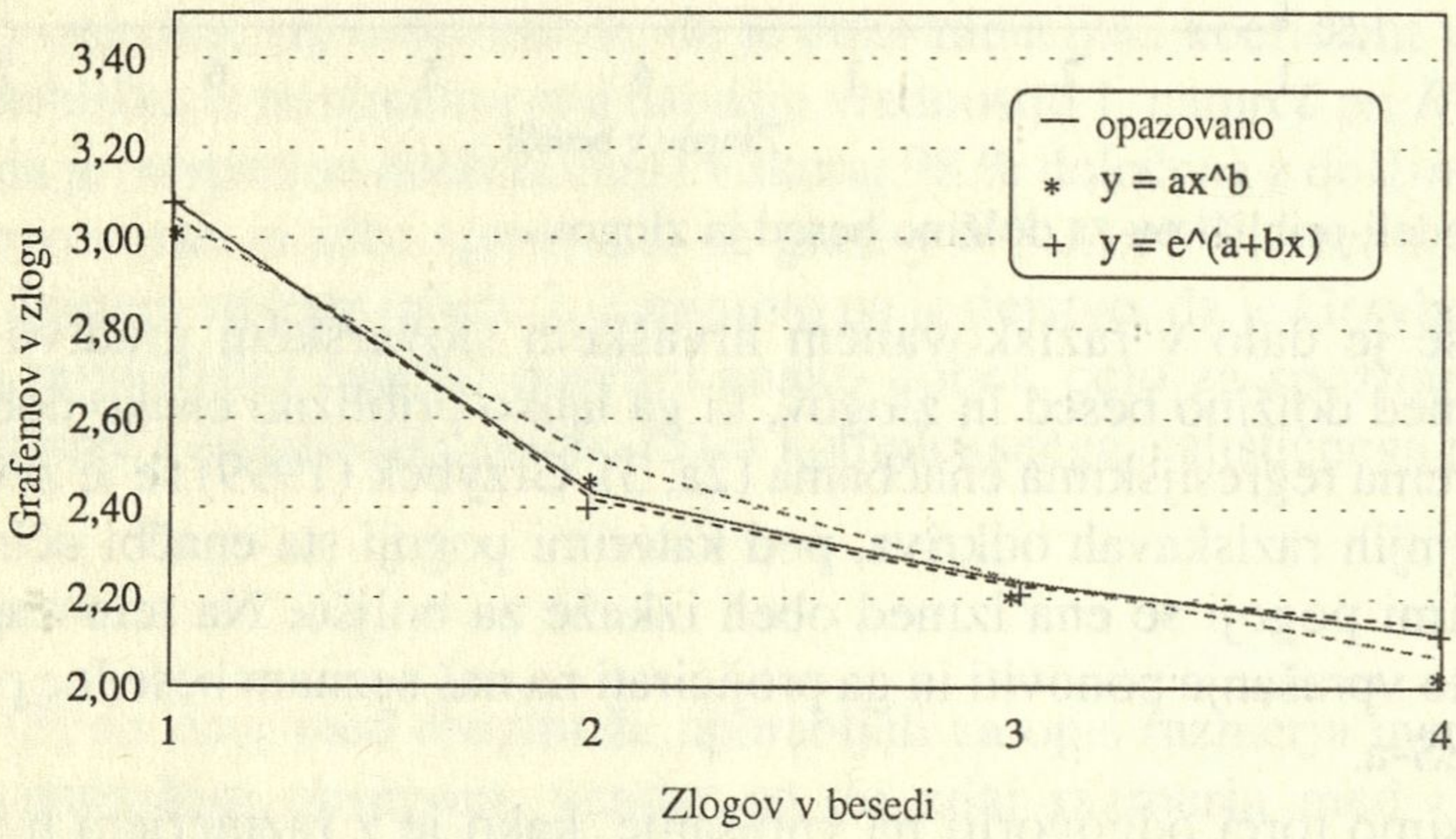
(4) $y = e^{(a+bx)}$

s parametroma $a = 0.64$ in $b = 0.49$ pri $R^2 = .998$ še boljši približek. Ustrezni podatki so v tabeli 3. Temeljijo na povprečnih vrednostih eno- do štirizložnih besed; pet »ničzložnih« besed v celotnem vzorcu ni bilo vštetih (z enim grafemom v zlogu ne izkazujejo variacije), pa tudi ne ena petzložnica.

Tabela 3: Odvisnost med dolžino besed in zlogov v besedni pogostnosti na podlagi CORTES-a

Zlogov v besedi	Grafemov v zlogu		
	opazovano	$y = ax^b$	$y = e^{(a+bx)}$
1	3,08	3,05	3,08
2	2,43	2,51	2,42
3	2,24	2,24	2,23
4	2,12	2,07	2,14
		$R^2 = .979$	$R^2 = .998$

Graf 3 prikazuje podatke in njihovo teoretično modeliranje.



Graf 3: Odvisnost med dolžino besed in zlogov v seznamu besedne pogostnosti na podlagi CORTES-a

Če povzamemo rezultate odvisnosti med dolžino besed in zlogov, vidimo, da sta pri slovenskem seznamu besedne pogostnosti prav tako kot pri hrvaškem slovarskem gradivu (obe raziskavi se ločita ne le po jeziku, ampak tudi po različnem upoštevanju besedne pogostnosti) različni regresijski enačbi približno enako dobri za teoretično modeliranje omenjenega razmerja.

Spričo teh opazovanj je posebej zanimivo, da se da obe formuli, kot predlagajo Wimmer/Köhler/Altmann (Ms.), izpeljati iz skupne, nadredne enačbe (5), pri čemer je ta enačba lahko podlaga za druge posebne primere, med drugim tudi za obe omenjeni enačbi:

$$(5) \qquad y = Ce^{a_0/x} x^{a_1} e^{-a_2/x - a_3/(2x^2) - a_4/(3x^3) - \ldots}$$

Za $a_1 < 0$, $a_0 = a_2 = a_3 = \ldots = 0$ dobimo zgoraj omenjeno formulo (II) Menzerath Altmannovega zakona:

$$(6) \qquad y = Cx^{a_1}.$$

In enačba

$$(7) \qquad y = e^{(a+bx)}$$

se da na predlog Wimmerja/Köhlerja/Altmanna (Ms.) v primeru $a_2 < 0$, $a_0 = a_1 = a_3 = ... = 0$, preoblikovati v:

(8) $y = Ce^{-a_2/x}$.

To peroblikovanje pripelje za Gajićeve podatke s parametroma $C = 1.89$ in $a = 0.49$ do enačbe (8a):

(8a) $y = 1.89e^{-0.49/x}$.

Vprašanje, ki ga zastavlja Grzybek (1999), v katerih okoliščinah sta enačbi enako učinkoviti, ostaja tako še vedno aktualno. Ponovno ga bo treba postaviti pri nadaljnjih raziskovanjih dolžine besed, pri katerih so pomembni še drugi dejavniki, ki vplivajo na dolžino, kot so dolžina stavkov in besedilni parametri. To vprašanje bo treba obdelati na drugem mestu, pri čemer bo treba ločevati med dvema različnima primeroma: na eni strani dolžino besed v besedilih – torej odvisnostjo med *stavčnimi* in *besedilnimi* dejavniki –, na drugi med dolžino besed v stavkih, npr. v pregovorih, pri katerih so verjetno pomembni *stavčni* in ne *besedilni* parametri (prim. Grzybek 2000a, b).

V nadaljevanju bomo to vprašanje pustili ob strani in se namesto tega posvetili vprašanju dolžine besed.

3. Dolžina besed

Kot smo že povedali, je za raziskavo dolžine besed izračunavanje srednjih vrednosti ne samo povsem navaden, ampak verjetno tudi najbolj uporabljani postopek. Zato začnimo našo analizo z vrsto splošnih trditev.

991 besednih oblik (*types*) se pojavlja 1.900.132 krat (*tokens*). Najpogostejša besedna oblika (na 1. mestu) se pojavlja s pogostnostjo *fi* = 172.035, najredkejša (na 991. mestu) z relativno še vedno visoko *fi* = 271. Povprečna dolžina besed, merjena v zlogih, je 2,025 ($s = 0,71$). V grafemih merjena povprečna dolžina besed znaša 4,922 (s = 1,58). Ta vrednost je znatno nižja od Jakopinove (1999) povprečne dolžine besed 4,55, ki jo je ugotovil na dveh različnih korpusih. Ker Jakopin ne navaja standardnih odklonov, ne moremo preveriti pomembnosti v razliki srednjih vrednosti med njegovimi in našimi vzorci. Vendar pa je predvsem ob upoštevanju obsežnosti vzorcev pri omenjenih srednjih vrednostih pričakovati veliko razliko. Razlog za to bi bil lahko – upoštevaje relativno varne domneve o (negativni) korelaciji med besedno dolžino in besedno pogostnostjo (»kolikor večja je besedna pogostnost, toliko krajša je beseda«) – da je Jakopin pri analizi upošteval ***vse*** besedne oblike v celotnem besedilnem korpusu, medtem ko zgornja analiza temelji na bolj ali manj poljubnem izboru 1000 (oz. 991) najpogostejših enot.

Na tem mestu pa se ne bomo zadovoljili s preprostim izračunavanjem srednje vrednosti in standardnega odklona, med drugim zato, ker srednje vrednosti včasih dajejo varljiv vtis, ker osvetlijo podatkovno gradivo le z določene perspektive. Tako so enake srednje vrednosti lahko rezultat različnih vrednosti (npr. $\bar{x} = 3,5$ za vzorec n_1 z vrednostmi 2,3,4,5 in $\bar{x} = 3,5$ za vzorec n_2 z vrednostmi 2,2,2,8) – to

razliko praviloma opišemo z varianco oz. standardnim odklonom (tj. s korenom iz variance) (v navedenem primeru $s_1 = 1.29$ proti $s_2 = 3.00$). Po drugi strani imamo lahko tudi delne vzorce z enakim standardnim odklonom na podlagi povsem različnih pogojev, kar bomo prav tako na kratko ponazorili z naslednjim izmišljenim primerom: vzemimo vzorec n_1 s pogostnostjo 1,2,3,4,5 in vzorec n_2 s pogostnostjo 3,4,5,6,7. Srednja vrednost v obeh vzorcih ni enaka, standardni odklon pa je. Če torej izhajamo iz tega, da je x_1 število enozložnih, x_2 pa dvozložnih besed itd., potem so vrednosti enake zato, ker je v enem vzorcu manj kratkih in več daljših, v drugem pa več kratkih in manj dolgih besed. To, kar se, z drugimi besedami, razlikuje v obeh vzorcih, je vrsta pogostnostne porazdelitve. Zato je smiselno, da poleg srednjih vrednosti in standardnih odklonov ne izračunavamo samo npr. drugih »mer centralne tendence« in razprševanja, kot to imenujejo v statistiki, ampak da raziskujemo tudi celotno obliko pogostnostne porazdelitve.

Glede na besedno pogostnost bi se morali v prvem koraku vprašati, kako pogosto se pojavljajo besede določene dolžine, v drugem koraku pa, ali lahko rezultate pogostnostne porazdelitve formaliziramo tudi teoretično in jih interpretiramo. Teh korakov se bomo lotili v 4. poglavju.

Še prej pa naj drugače opozorimo na problem porazdelitve besednih pogostnosti in na nujnost njihove raziskave: če namreč na podlagi zgoraj omenjenega izhajamo iz tega, da so povprečne dolžine odvisne od vseh besed v korpusu, ne odgovorimo na vprašanje, ali so bolj ali manj dolge besede v korpusu porazdeljene enakomerno. Na podlagi dobro utemeljene domneve o korelaciji med besedno dolžino in pogostnostjo bi bilo pričakovati, da je povprečna dolžina pri pogostejših besedah (ki so torej v padajočem zaporedju na višjih mestih) krajša od dolžine manj pogostih.[3] Preden se lotimo teoretičnega modeliranja pogostnostne porazdelitve, si torej poglejmo še to vprašanje v zvezi z seznamom CORTES.

Če iz omenjenega razloga celotni vzorec ločimo na pet približno enakih delnih vzorcev (štiri po 200, eden po 186 enot) in izračunamo srednjo besedno dolžino za vsak delni vzorec, je to v določenem smislu seveda poljubna razmejitev. Ne glede na to se potrdi naše pričakovanje, in sicer tako pri izračunu povprečne dolžine zlogov v besedi kot tudi grafemov v besedi: kot se vidi iz 2. in 5. stolpca v tabeli 4, narašča povprečno število zlogov v besedi od prvega do petega vzorca od $\bar{x} = 1{,}54$ do $\bar{x} = 2{,}27$, povprečno število grafemov v besedi od $\bar{x} = 3{,}80$ do $\bar{x} = 5{,}45$. Zanimivo je, da je ta težnja enako močno izražena pri merjenju besedne dolžine v zlogih in v grafemih – oba računska postopka dejansko zelo visoko korelirata ($r = .99$, $p < 0.001$).

Še bolj zanimivo pa je dejstvo, da tudi ta trend (daljših zlogov pri manj pogostih besedah), kot kažejo ustrezne regresijske analize, lahko formaliziramo. Pri tem pa z Menzerath-Altmannovo formulo (II) izračunamo le šibke približke: Pri izračunu povprečnih besednih dolžin v zlogih znaša $R^2 = .8857$ in tudi pri izračunu v grafemih pridemo le na $R^2 = .8715$. Izrazito boljši pa je približek z regresijsko enačbo (8) $y = Ce^{-a/x}$. V tem primeru dosežejo približki po vstavitvi $C = 2.510$ in $a =$

[3] Pri izračunu povprečnih besednih dolžin je bilo iz korpusa izločenih še 5 enot, in sicer t. i. ničzložnic, ki so sestavljene samo iz enega nezlogotvornega soglasnika. Obseg raziskovanega korpusa znaša potemtakem $N = 986$.

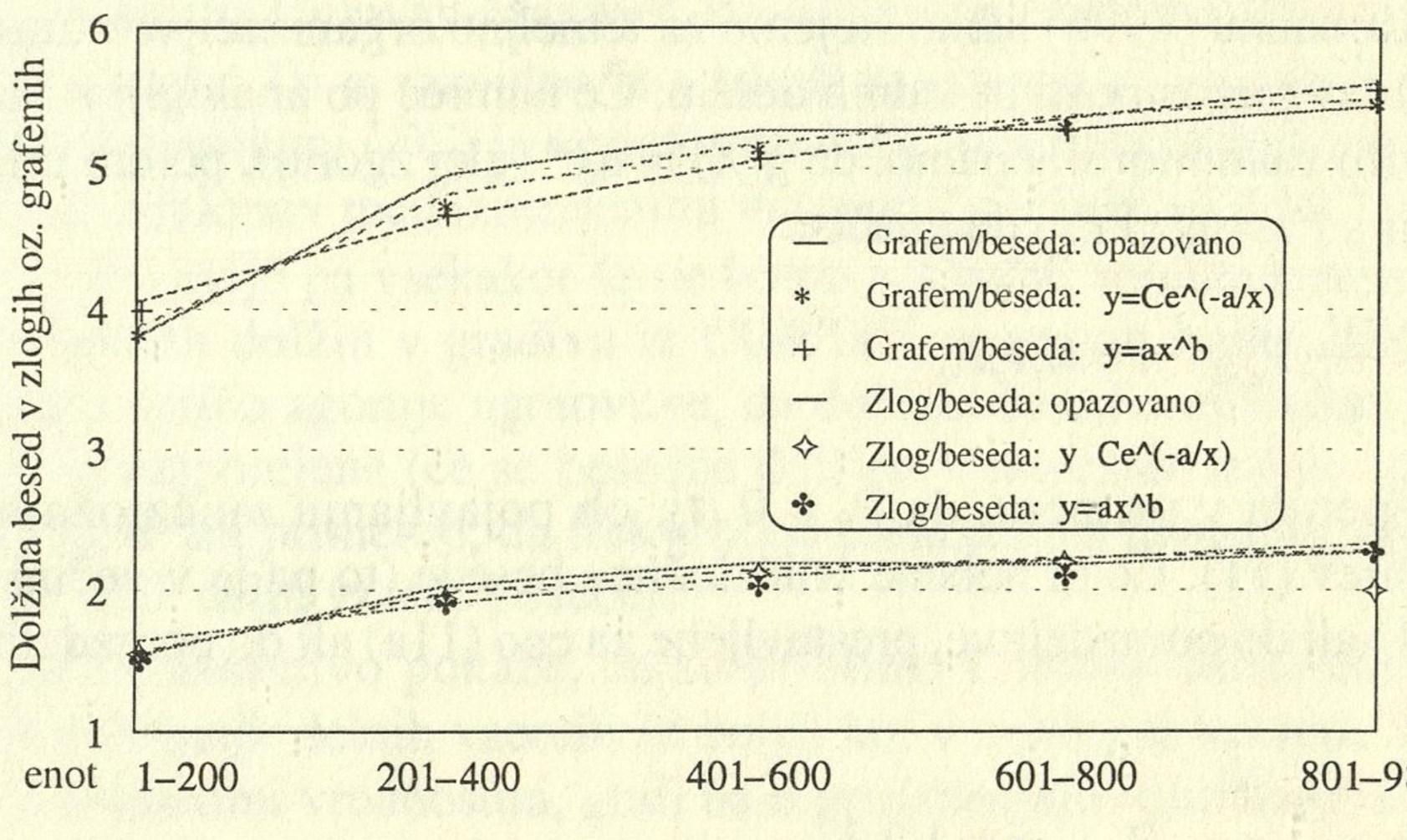
-0.475 oz. $C = 6.002$ in $a = -0.439$ v obeh primerih (to pomeni pri izračunu besedne dolžine na podlagi števila zlogov oz. grafemov v besedi) visok determinacijski koeficient $R^2 = .98$.

Rezultat lahko pojasnimo v tem smislu, da se na isti način očitno ne da formalizirati le razmerja med dolžino besed in zlogov (glej zgoraj), temveč tudi razmerje med besedno dolžino in besedno pogostnostjo. Če na podlagi te ugotovitve postavimo obratno hipotezo, bi mogli reči, da je v obeh primerih to razmerje posledica iste zakonitosti. Vprašanja, ali se z regresijsko enačbo (8), ki modelira to razmerje, da doseči dober približek tudi takrat, ko je na seznamu besedne pogostnosti več kot 1000 izbranih najpogostejših enot in če oz. kako se da to utemeljiti, se bo treba lotiti na drugem mestu. – Tabela 4 vsebuje izmerjene in teoretično izračunane vrednosti.

Tabela 4: Naraščanje besedne dolžine pri manj pogostih besedah

Enot	Zlogov v besedi			Grafemov v besedi		
	(opazovano)	$y = Cx^a$	$y = Ce^{(-a/x)}$	(opazovano)	$y = Cx^a$	$y = Ce^{(-a/x)}$
1-200	1,54	1,64	1,56	3,80	4,05	3,87
201-400	2,03	1,91	1,98	4,96	4,66	4,82
401-600	2,19	2,09	2,14	5,27	5,06	5,18
601-800	2,18	2,22	2,23	5,28	5,36	5,38
801-986	2,27	2,33	2,28	5,45	5,61	5,50
		$R^2 = .88$	$R^2 = .98$		$R^2 = .87$	$R^2 = .98$

Graf 4 prikazuje razmerje med besedno dolžino in besedno frekvenco.



Graf 4: Naraščanje besedne dolžine ob padajoči pogostnosti

Zdaj se lahko posvetimo četrtemu sklopu vprašanj, pogostnostni porazdelitvi besednih dolžin in njenemu teoretičnemu modeliranju.

4. Pogostnost besednih dolžin in njena porazdelitev

Kot smo že omenili, gre pri analizi pogostnosti besednih dolžin, prvič, za vprašanje, kako pogosto se v korpusu pojavljajo besede določene dolžine (to pomeni,

koliko je eno-, dvo-, trizložnih itd. besed), in drugič, ali lahko rezultat pogostnosti porazdelitve teoretično formaliziramo in modeliramo.

Temeljna domneva pri modeliranju pogostnostne porazdelitve besednih dolžin je, da razredi dolžin – na podlagi besedil ali slovarjev – ne kažejo kaotične porazdelitve, ampak so v določeni sorazmernosti. Ta odnos, ki ga v splošni obliki lahko razumemo kot

(9) $P_x \sim P_{x-1}$

pravi, da je obseg vsakega »višjega« razreda odvisen od prejšnjega »nižjega« razreda. To sorazmernost lahko opišemo s funkcijo $g(x)$, s tem izpeljemo naslednjo splošno trditev:

(10) $P_x = g(x)\, P_{x-1}$.

Za $g(x)$ poznamo celo vrsto različnih funkcij in glede na vrsto funkcije tudi različne porazdelitvene modele (prim. Wimmer et al. 1994, Wimmer/Altmann 1996). Natančnejši predstavitvi in razpravi o teh modelih se lahko tukaj odrečemo in se takoj posvetimo rezultatu naše analize seznama besedne pogostnosti na podlagi CORTES-a. Najbolje se je t. i. Conway-Maxwell-Poissonovim porazdelitev pokrivala z empiričnimi podatki po naslednji formuli:

(11) $$P_x = \frac{a^x}{(x!)^b} P_0 \quad x = 0,1,2,...$$

Ta rezultat je zanimiv zato, ker Conway-Maxwell-Poissonov približek po Wimmer/Altmannu (1996) lahko štejemo za temeljno organizacijsko funkcijo $g(x)$ za porazdelitev pogostnosti besednih dolžin: Če namreč po analogiji z Menzerath-Altmannovim zakonom sklepamo, da $g(X) = ax^{-b}$ (glej zgoraj), potem pripelje to v kombinaciji s $P_x = g(x)\, P_{x-1}$ do enačbe:

(11a) $$P_x = \frac{a}{x^b} P_{x-1} \quad x = 0,1,2,...$$

To pa spet da v primeru, da $P_0 \neq 0$ (tj. ob pojavljanju »ničzložnih« besed) zgornjo rešitev (11). Če ni nobene »ničzložne« besede (to pa je v večini jezikov), pripelje (11) ali do porazdelitve, prestavljene za eno (11a) ali do porazdelitve, oprte na 0 (11b):

(11a) $$P_x = \frac{a^{x-1}}{(x-1)!^b} P_{x-1} \quad x = 1,2,3,...$$

(11b) $$P_x = \frac{a^x}{(x!)^b} P_1 \quad x = 1,2,3,...$$

V tabeli 5 so ustrezni podatki: V prvem stolpcu (x) je vsakokratno število zlogov v besedi, v drugem stolpcu (fx) empirično izmerjeno število vsakokratnih x-zložnih besed iz CORTES-ovega seznama besedne pogostnosti, v tretjem

stolpcu (*NPx*) pa so navedeni teoretično izračunani rezultati na podlagi ujemanja s Conway-Maxwell-Poissonovo porazdelitvijo. V spodnjih vrsticah tabele 5 so vrednosti parametrov *a* in *b*, χ^2-vrednost χ^2-testa z ustreznimi prostostnimi stopnjami (*FG*), s pripadajočo verjetnostjo *P* in s kontingenčnim koeficientom *C*.

Rezultati porazdelitvenega modeliranja za celotni vzorec so prikazani na grafu 5, ki spada k tabeli 5.

Zlogov v besedi		
x	*fx*	*NPx*
0	5	5,10
1	205	232,53
2	564	525,13
3	195	204,41
4	21	22,85
5	1	0,99
a	45,61	
b	4,34	
χ^2	6,71	
FG	2	
P	0,03	
C	0,0068	

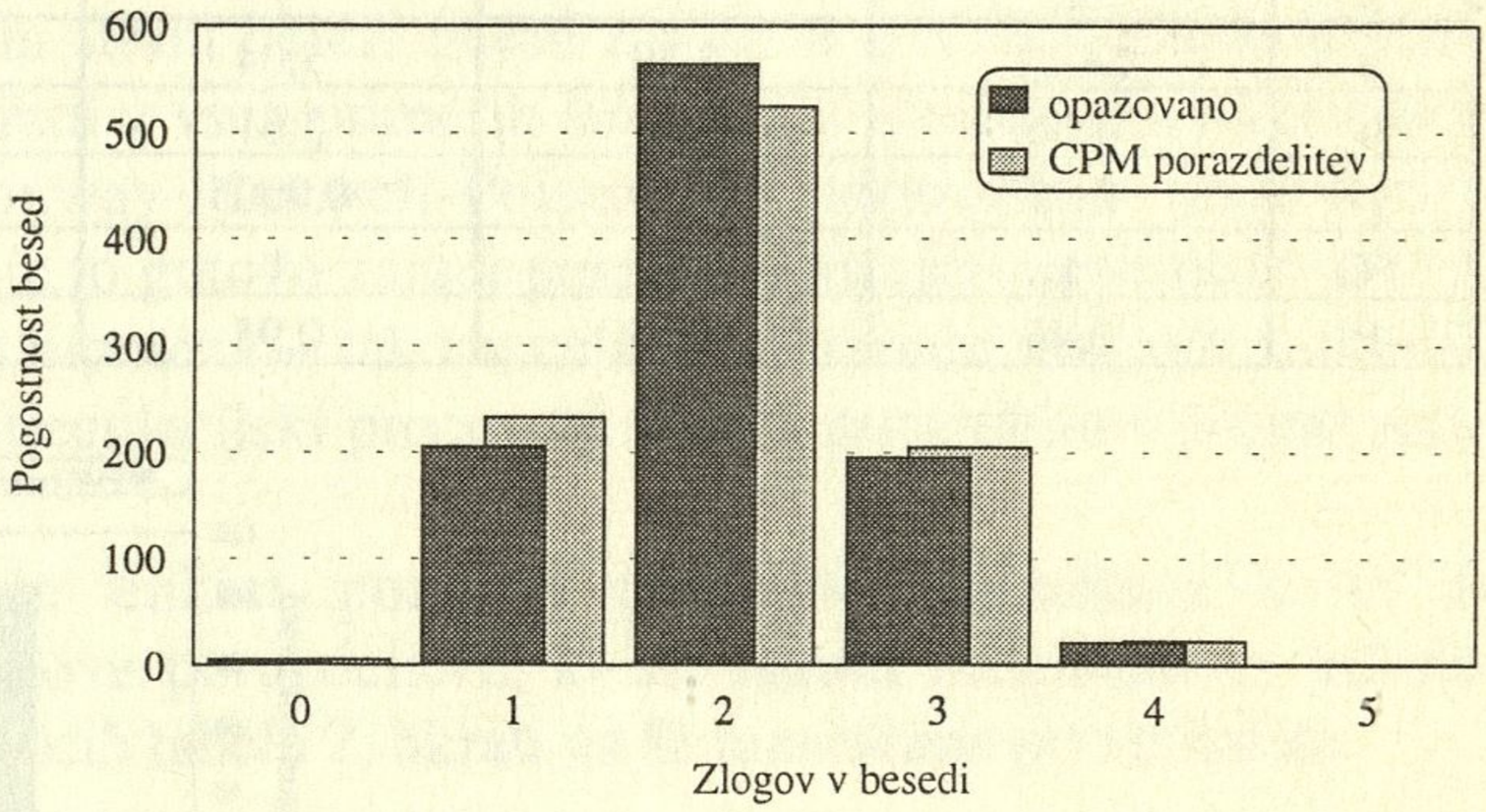


Tabela 5/Graf 5: Ujemanje Conway-Maxwell-Poissonove porazdelitve s celotnim vzorcem

Kot je videti, se Conway-Maxwell-Poissonova porazdelitev dejansko izkaže za odličen približek.[4] To je razvidno iz v tabeli navedenih χ^2-vrednosti z ustreznimi prostostnimi stopnjami (*FG*) in verjetnostmi (*P*), ki v nobenem primeru ne kažejo pomembnih odklonov med izmerjenimi in teoretičnimi vrednostmi.[5]

S to ugotovitvijo pa vsekakor še ne bomo zaključili analize frekvenčne porazdelitve besednih dolžin v gradivu iz CORTES-a, ampak bomo storili še korak naprej. Kajti spričo zgornje ugotovitve, da dolžine besed v celotnem vzorcu niso enakomerno razporejene (če se besedne dolžine v ustreznih delnih vzorcih jasno razlikujejo), se zdi primerno, da frekvenčno porazdelitev besednih dolžin raziščemo tudi za vsak delni vzorec posebej.

Pri tem se zanimivo pokaže, da so približki Conway-Maxwell-Poissonove porazdelitve v petih delnih vzorcih še boljši kot v celotnem vzorcu. V tabeli 6 so podatki z ustreznimi vrednostmi, grafi 6a-e predstavljajo rezultate grafično.

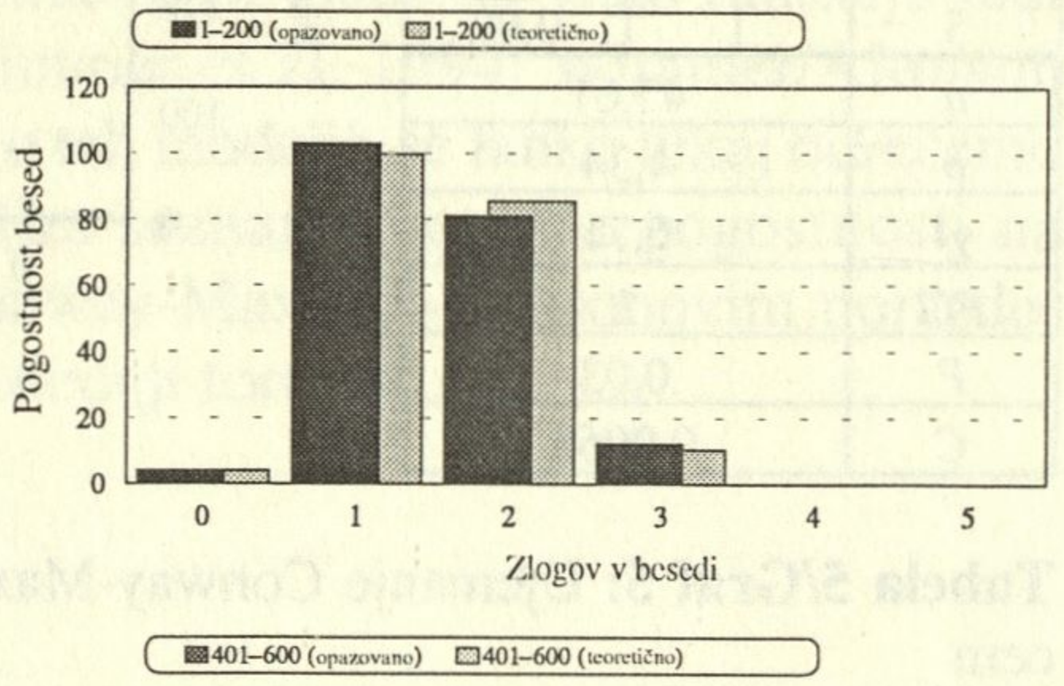
---

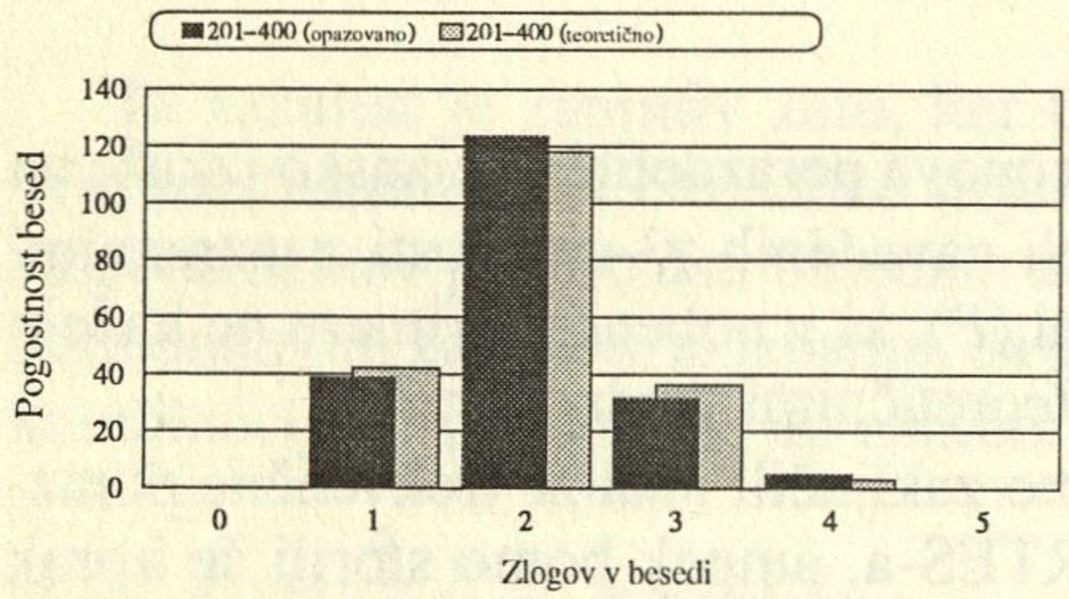
[4] V enem samem primeru so bili podatki združeni zaradi boljšega približka, in sicer v delnem vzorcu 601 – 800. Tukaj je bila ena ničzložna beseda prišteta k enozložnim. Ustrezna vrednost je v tabeli označena (|).

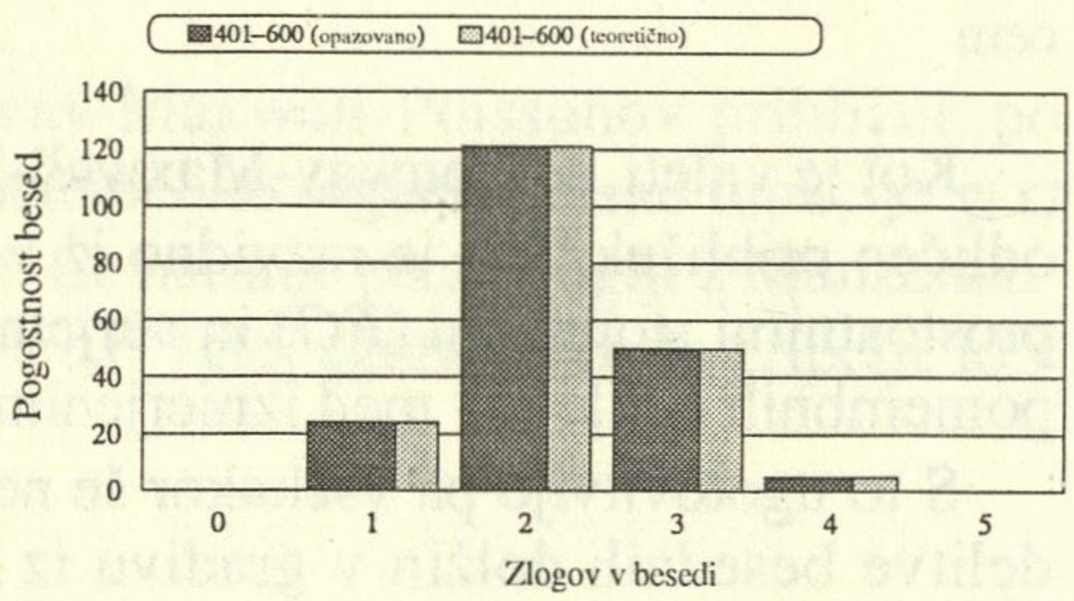
[5] Razlike med izmerjenimi in teoretičnimi vrednostmi štejejo za pomembne oz. zelo pomembne v primeru $P(\chi^2) < 0.05$ oz. $P(\chi^2) > 0.05$. Samo za celotni vzorec je dodatno naveden še kontingenčni koeficient, ker tukaj χ^2-vrednost (domnevno zaradi velikosti vzorca) v primerjavi z danimi vzorci kaže malenkostno večje odklone.

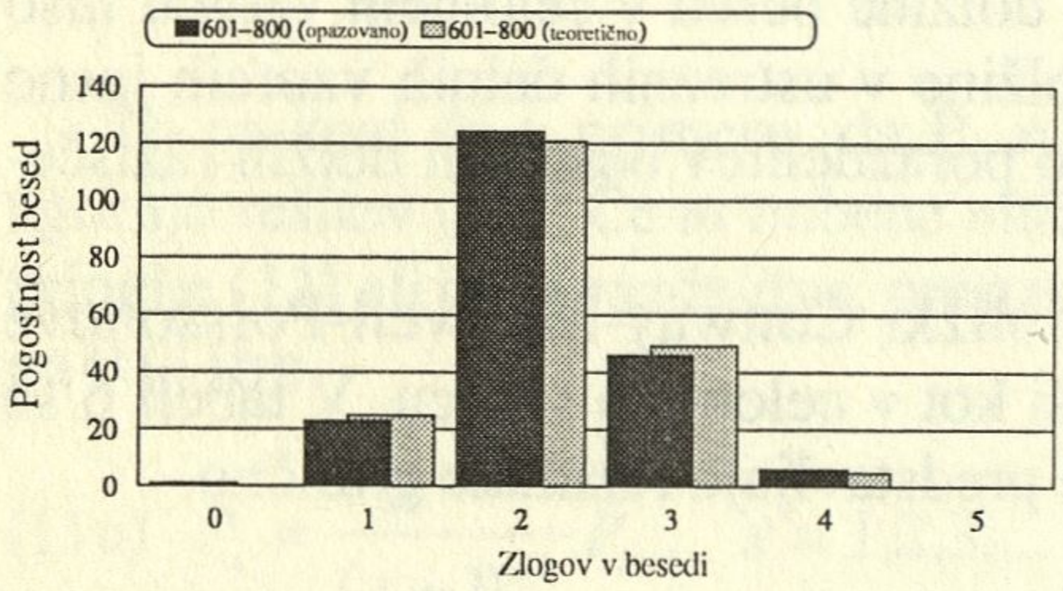
Tabela 6: Ujemanje Conway-Maxwell-Poissonove porazdelitve z delnimi vzorci

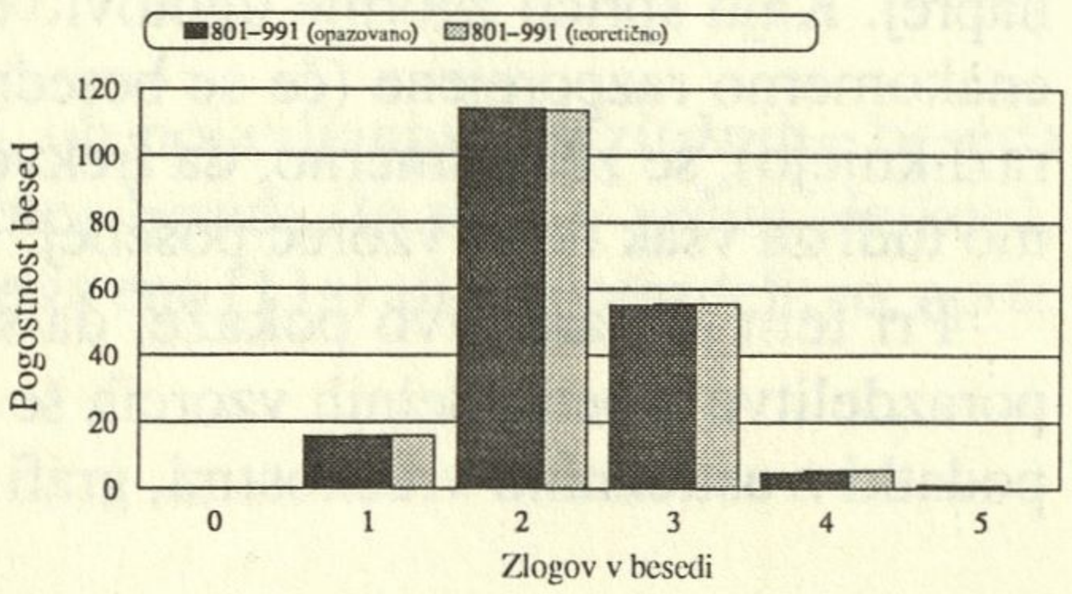
	1–200		201–400		401–600		601–800		800–991	
x	*fx*	*NPx*	*fx*	*NPx*	*fx*	*NPx*	*fx*	*NPx*	*fx*	*NPx*
0	4	4,02					1			
1	103	99,94	39	42,37	24	24,00	23	24,87	16	16,06
2	81	85,54	124	117,99	121	121,00	124	120,96	114	113,59
3	12	10,51	32	36,42	50	50,00	46	49,30	55	55,48
4			5	3,21	5	4,95	6	4,87	5	5,67
5									1	0,19
a	24,87		2,79		5,04		4,86		7,07	
b	4,86		3,17		3,61		3,58		3,86	
χ^2	0,55		2,10		0,0005		0,5873		11,22	
FG	1		1		1		1		1	
P	0,46		0,15		0,98		0,44		0,93	











Grafi 6a-e: Približki Conway-Maxwell-Poissonovi porazdelitvi v delnih vzorcih

Če povzamemo rezultate naše analize porazdelitve pogostnosti besednih dolžin, lahko upravičeno trdimo, da Menzerath-Altmannov zakon v našem seznamu besednih pogostnosti na podlagi CORTES-a določa porazdelitev pogostnosti besednih dolžin. Podatki v tabeli 6 oz. grafi 6a-e ob tem jasno kažejo, da odlični približek Conway-Maxwell-Poissonovi porazdelitvi v vseh petih delnih

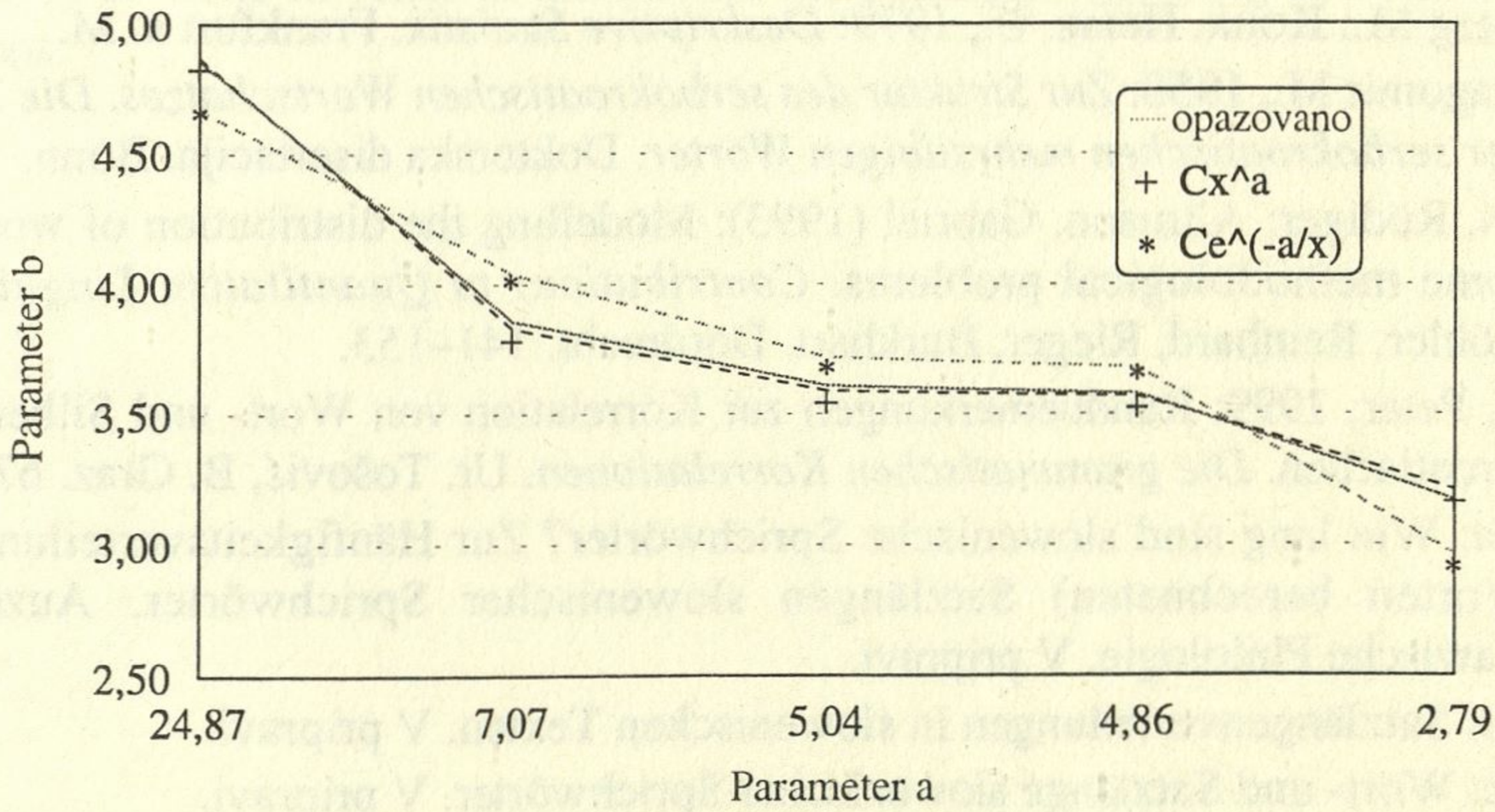
vzorcih nikakor ne implicira tudi ustreznega odstotnega deleža *x*-zložnih besed: Tako npr. jasno vidimo, da v delnem vzorcu prvih 200 besednih oblik prevladujejo enozložne besede, da pa sploh ni štirizložnih, medtem ko v ostalih štirih vzorcih večji delež predstavljajo dvozložne besede. Prav tako na primer vidimo, da je delež trizložnih besed največji v petem vzorcu.

Iz teh opažanj – ki se mimogrede pokrivajo z zgoraj predstavljenim dejstvom, da dolžine naraščajo ob padajoči pogostnosti, ki pa ga potrjujejo iz povsem drugega vidika – lahko na koncu izpeljemo ne nezanimivo hipotezo: če se namreč v teh delnih vzorcih po eni strani deleži besed z *x* oz. *x* + *1* itd. zlogov močno razhajajo in če po drugi strani v vseh primerih (to pomeni v celotnem vzorcu in v petih delnih vzorcih) Conway-Maxwell-Poissonova porazdelitev predstavlja najboljši približek, potem je to gotovo zaradi posebne variacije parametrov *a* in *b* te porazdelitve, ki – tako hipoteza – tako v celotnem vzorcu kot tudi v delnih vzorcih uravnavata (samo)regulacijski proces, ki določa porazdelitev pogostnosti besednih dolžin.

Ob tem vprašanju še enkrat pozorno poglejmo parametra *a* in *b* Conway-Maxwell-Poissonove porazdelitve, ki sta zaradi preglednosti ločeno navedena v prvih dveh stolpcih tabele 7, hkrati pa še razvrščena po velikosti.

Tabela 7: Parametra *a* in *b* v Conway-Maxwell-Poissonovi porazdelitvi

Vzorec	*a*	*b*	Cx^a	$Ce^{(-a/x)}$
1–200	24,87	4,86	4,87	4,67
201–400	7,07	3,86	3,83	4,04
401–600	5,04	3,61	3,59	3,72
601–800	4,86	3,58	3,57	3,68
801–991	2,79	3,17	3,21	2,96
			$R^2 = .9986$	$R^2 = .9119$



Graf 7: Razmerje med parametroma *a* in *b* v Conway-Maxwell-Poissonovi porazdelitvi

Kot je videti, se s padajočim *a* zmanjšuje tudi *b* (oz. *a* s padajočim *b*). Ta odvisnost pa nikakor ni, kot kažejo analize, linearna, ampak ustreza Menze-

rath-Altmannovemu zakonu po formuli (II) $y = ax^b$ oz. (6) $y = Cx^a$. Tabela 7 kaže te teoretične vrednosti, ki z R^2 v vrednosti .9986 dosegajo odličen približek. Jasno se tudi vidi, da je približek v korelaciji med *a* in *b* po enačbi $y = Ce^{(-a/x)}$ ob R^2 v vrednosti .9119 v tem primeru bistveno slabši. Graf 7 ponazarja to razmerje.

5. Sklep

S temi ugotovitvami zaključimo našo analizo dolžin in pogostnosti 1000 najpogostejših besed iz korpusa CORTES. Razvidno je, da pogostnost in dolžina besed, kakor tudi njuni porazdelitvi nista naključno organizirani jezikovni količini, ampak da korelirata z različnimi drugimi dejavniki, ki skupaj predstavljajo samoregulacijski krog. Obravnavali smo lahko le nekaj teh dejavnikov, med drugim zato, ker nimamo opraviti z besedami v besedilu, ampak z besednim seznamom, in ker gre povrh vsega pri seznamu besedne pogostnosti dejansko za rezultat raznovrstnega besedila. Na vprašanje, ali se dajo opazovani procesi modelirati na podoben ali drugačen način v besedilih in katere dodatne parametre je treba upoštevati, bo treba odgovoriti na drugem mestu (Grzybek 2000a,b). Še precej dela nas čaka, da bomo doumeli zakone, ki vplivajo na zgoraj prikazane korelacije.

Literatura

Altmann, Gabriel, 1980: Prolegomena to Menzerath's Law. *Glottometrika 2*. Ur. Grotjahn, Rüdiger. Bochum. 1–10.

– – 1992: Das Problem der Datenhomogenität. *Glottometrika 13*. Ur. Rieger, Burghard. Bochum. 287–298.

Altmann, Gabriel, Schwibbe, Michael H., 1989: *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim u. a.

Bunge, Mario, 1967: *Scientific Research II: The Search for Truth*. New York.

Diehl, Joerg M.; Kohr, Heinz. U., 1979: *Deskriptive Statistik*. Frankfurt a. M.

Gajić, Dragomir M., 1950: *Zur Struktur des serbokroatischen Wortschatzes. Die Typologie der serbokroatischen mehrsilbigen Wörter*. Doktorska disertacija. Bonn.

Grotjahn, Rüdiger; Altmann, Gabriel (1993): Modelling the distribution of word length: some methodological problems. *Contributions to Quantitative Linguistics*. Ur. Köhler, Reinhard, Rieger, Burkhart. Dordrecht. 141–153.

Grzybek, Peter, 1999: Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen. *Die grammatischen Korrelationen*. Ur. Tošović, B. Graz. 67–77.

– – 2000a: Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter. Auzeiger für slawische Philologie. V pripravi.

– – 2000b: Satzlängenverteilungen in slowenischen Texten. V pripravi.

– – 2000c: Wort- und Satzlänge slowenischer Sprichwörter. V pripravi.

Jakopin, Primož, 1996: Ali so rojstna imena krajša od drugih samostalnikov? *Slavistična revija* 44/2. 193–200.

– – 1999: *Zgornja meja entropije pri leposlovnih besedilih v slovenskem jeziku*. Doktorska disertacija, Ljubljana. [http://www.ff.uni-lj.si/pj/disertacija]

NEMCOVÁ, Emilia, ALTMANN, Gabriel, 1994: Zur Wortlänge in slowakischen Texten. *zet – Zeitschrift für empirische Textforschung* 1. 40–43.

ORLOV, Jurij K., 1982: Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie 'Sprache – Rede' in der statistischen Linguistik). *Sprache, Text, Kunst. Quantitative Analysen*. Ur. Orlov, Ju. K., Boroda, M. G., Nadarešvili, I. Š. Bochum. 1–55.

SACHS, Lothar, 1992: *Angewandte Statistik. Anwendung statistischer Methoden*. Heidelberg u. a.

WIMMER, Gejza, ALTMANN, Gabriel, 1996: The Theory of Word Length: Some Results and Generalizations. *Glottometrika 15*. Ur. Schmidt, Peter. Trier. 112–133.

WIMMER, Gejza, KÖHLER, Reinhard, ALTMANN, Gabriel, 2000: Unified derivation of some linguistic laws. *Handbook of Quantitative Linguistics*. V pripravi.

WIMMER, Gejza, KÖHLER, Reinhard, GROTJAHN, Rüdiger, Altmann, Gabriel, 1994: Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics* 1/1. 98–106.

Iz nemščine prevedel
Darko Čuden.

SUMMARY

The author demonstrated in the paper that frequency and length of lexical items as well as their distribution are not coincidentally organized language quantities, but, rather, they are related to other factors, all of which in combination represent a self-regulatory circle. Only a few of these factors could be discussed in the article, partially because the analyzed lexicon was from a glossary rather than from a text, and because the list of lexical frequency was derived from a variegated text. The questions of whether the analyzed processes can be modeled using the same or different method and in texts, and what additional parameters need to be taken into account, will have to be answered on another occasion. Much work still needs to be done in order to understand the principles affecting the aforementioned correlations.