Zapiski, ocene in poročila

ČEŠKI FREKVENČNI SLOVAR

Kvantitativno raziskovanje jezika je še zelo mlada jezikoslovna veja, ki se je začela razvijati z uveljavljanjem matematične lingvistike, strojnega prevajanja in teorije informacije. Omenjena področja jezikoslovja se do neke mere med seboj prepletajo, vsem pa je skupna naslonitev na številčno podajanje pri raziskavah jezika. Frekvenca besede nam s številčnim podatkom kaže njeno pogostnost nastopanja v jeziku, plásti jezika ali nekega dela. Proučevanje frekvence besed pa ne služi zgolj raziskovanju odnosov, ki vladajo v jeziku, ampak dobi npr. neposredno praktičen pomen kot pomoč pri izdelavi dvojezičnih slovarjev, posebno pa še pri izdelavi t. im. minimalnih slovarjev za učenje tujih jezikov.

Frekvenčni slovarji so od prvih poskusov do danes zadevali ob različno problematiko v jeziku in služili različnim ciljem. Češka raziskovalka frekvence besed Marie Těšitelová vidi pomembnost frekvenčnih slovarjev posebno za 1. spoznavanje leksikalne strani jezika, 2. reševanje nekaterih stilističnih vprašanj in 3. za prispevanje k spoznavanju nekaterih gramatičnih vprašanj, posebno v morfologiji (SaS, XXII, 1961, str. 173). Raziskovanje frekvence besed pa ima praktično vrednost tudi pri sestavljanju učbenikov.

Frekvenca besed raste ali pada z njihovo aktualnostjo in je torej lahko v različnih dobah in stilnih področjih različna. Idealno bi bilo imeti frekvenčni slovar vsake generacije. Tako bi brez težav sledili frekvenco besed skozi zgodovinski razvoj jezika in ugotavljali njihove vsakokratne funkcije in selitve besed iz ene jezikovne plasti v drugo. Pomagali bi nam pri proučevanju raznih jezikovnih zakonitosti itd.

Frekvenčni slovar besed, besednih vrst in oblik v češkem jeziku[1] naravnost ne zadeva nobenega od omenjenih področij moderne lingvistike, vendar pa je trdna osnova nadaljnjemu kvantitativnemu študiju češkega jezika. Iz takšnega, kakršen je sedaj, pa lahko črpajo dragoceno gradivo vsa področja. Izredno bogat vir za raziskovanje pa ni samo češkemu lingvistu in pedagogu, ampak tudi raziskovalcu kvantitativnih razmerij katerega koli jezika.[2] Iz tega razloga in pa zato, ker se je v zadnjem času tudi pri nas razpravljalo o tem problemu, kratka informacija o češkem frekvenčnem slovarju ne more biti odveč.

Avtorji so si že ob začetku svojega dela zastavili zelo zahtevne naloge. Niso hoteli podati samo kvantitativnega stanja v jeziku, temveč so dobljeni številčni material uspešno dopolnili s kvalitativno analizo. Sami poudarjajo, da bi bila velika napaka, če bi številkam, ki jih dobimo s kvantitativno analizo, pripisovali končno veljavnost; svojo tehtnost dobijo namreč šele tedaj, če jih dopolnjuje kvalitativna analiza. To seveda ne pomeni, da mora biti vsaka številčna analiza združena s kvalitativno; če ni, je treba nanjo gledati kot na material (str. 9). Poleg tega, da prinaša slovar bogat številčni material, pa je njegova dobra stran prav v uspešni dopolnitvi kvantitativne analize s kvalitativno, kar omogoča koristne zaključke, npr. glede stilne vrednosti besed, frekvence pasiva v različnih rabah jezika itd.

V slovarju je poudarjeno, da se je treba pri vrednotenju posameznih frekvenc (frekvenc v posameznih virih) ozirati tudi na dobo, v kateri je ekscerpirano delo nastalo. Kar zadeva dobo, je iz seznama ekscerpiranih del razvidno, da so najmočneje — z nad 50 % — zastopane letnice 1946, 1947 in 1948 (najstarejša letnica je 1926), kar potrjuje, da se avtorji pri izbiri virov za ekscerpiranje niso ravnali po vnaprej določenem seznamu, ampak so ga sproti dopolnjevali z najnovejšimi viri. Tako je slovar zanimiv tudi s te strani, saj statistično zajema jezikovno stanje obdobja, ki je zelo pomembno, v jeziku pa izredno pestro.

[1] Jaroslav Jelínek, Josef V. Bečka, Marie Těšitelová: Frekvence slov, slovních druhů a tvarů v českém jazyce, Praha 1961. 586 str.

[2] Tako npr. J. V. Bečka primerja frekvenco besednih vrst v nekaterih funkcijskih stilih v češčini in francoščini. (Český jazyk a literatura, XIV, 224. sl.). Njegovo zanimivo ugotovitev, da so razlike v posameznih funkcijskih stilih obeh jezikov praviloma zelo majhne in da so razlike med posameznimi funkcijskimi stili istega jezika večje, je potrdil tudi M. Renský na osnovi primerjave frekvenc besednih vrst v angleščini (ČJL, XIV, 422 sl.) .

Gradivo, iz katerega je slovar nastal, tvori 1,623.527 besed, dobljenih s popolnimi izpisi petinsedemdesetih del. Pri izbiri so upoštevani različni jezikovni stili, po katerih je gradivo razvrščeno v osem skupin: leposlovje (A), poezija (B), mladinska literatura (C), dramatika (D), strokovna literatura (E), žurnalistika (F), znanstvena literatura (G), govorjeni jezik (H).[3] Slovar je bil od vsega začetka zasnovan tako, da je bilo gradivo mogoče porabiti ne samo za določitev čiste frekvence, ampak tudi za ugotavljanje, kako so zastopani posamezni sklanjatveni tipi (pri vsakem geslu se je že pri ekscerpiranju označeval tip sklanjatve), kakšna je frekvenca sklonov, spola in števila pri sklanjatvi in kakšna je frekvenca glagolskih oblik.

Orientacijsko bi lahko slovar razdelili na dva dela. Prvi del prinaša bogat pregled in oznako podobnih dosedanjih del iz svetovne lingvistike, govori o principih, po katerih so izbirali dela za ekscerpiranje, ter o metodi in tehniki dela, predvsem pa komentira dobljeno številčno gradivo in izvaja iz njega splošne zaključke. Tu je tudi cela vrsta tabel in diagramov, ki sistematično in nazorno razvrščajo dobljeno gradivo ter omogočajo zanimive primerjave in vrednotenje frekvence besed.

Drugi del je slovar frekvence besed. Obsega dva seznama. Seznam besed po frekvenci je sestavljen tako, da se začenja z besedo najvišje frekvence, navedenih pa je od 54.486 besed samo prvih 10.000, frekvenco pa kažejo tri številke. Pri razvrščanju je odločala absolutna frekvenca, ugotovljena v materialu (prva številka), število skupin, v katerih je beseda nastopila (druga številka), in število virov, v katerih je beseda izpričana (tretja številka). Besede z enako absolutno frekvenco so razvrščene glede na večje število virov, pri enaki frekvenci in enakem številu virov pa je beseda uvrščena na višje mesto, če nastopa v večjem številu skupin. Pri enakih vseh treh številkah odloča o razvrstitvi abeceda.

Najmočnejšo frekvenco ima v češčini veznik *in* (češko *a*). Nastopa v vseh osmih skupinah in vseh 75 delih 67.122 krat. Za njim se zvrstijo glagol *biti* (sem), *ta, v, on, na, da* itd., srednjo frekvenco ima npr. *vozel* (32-5-15), *upirati (se)* (31-7-21), *kratek* (27-6-17), nizko frekvenco pa npr. *porod* (14-6-08), *ponižati (se)* (14-5-08), *rdečelas* (14-2-07).

Podatke prvega seznama pa bistveno dopolnjuje drugi, abecedni seznam, ki se začenja s frekvenco 3. Ta seznam označuje frekvenco po posameznih skupinah, to se pravi, pove nam, kako je beseda zastopana v različnih stilih. Če smo torej iz prvega seznama razbrali absolutno frekvenco, število skupin in virov, v katerih se je beseda pojavila, nam drugi seznam ta podatek podrobneje razčleni in pokaže, v katerih skupinah se je beseda pojavila, kolikokrat v posamezni skupini ter v koliko knjigah:

	A	B	C	D	E	F	G	H
človek	2705-8-71:	982/15	61/10	269/10	377/10	246/6	89/7	531/9	141/4
tolažba	43-6-23:	18/7	6/3	10/5	6/6			1/1	2/1
motor	701-6-17:	16/2	6/4	11/1		638/2	10/1	20/4	
hrepeneti	5-3-5 :	2/2		2/2	1/1				
nazor	311-7-45:	35/13		8/4	4/4	71/7	34/6	130/7	29/4
negativen	37-2-7 :					5/1		22/6	
korelacija								71/4	

Izbrani primeri nam mnogo povedo: besedi *človek, motor* imata visoko frekvenco, spadata torej v osnovni besedni fond vsakdanjega jezika, nastopata skoraj v vseh skupinah (stilnih področjih), vendar sta najmočneje zastopani v skupinah ABCD, torej v manj strokovnih skupinah, razen druge, ki ima izjemno visoko frekvenco v skupini E, v strokovnem tekstu (K. Chochola, Izgorevalni motorji). Beseda *tolažba* je najbolj pogostna v leposlovju, kjer se pojavi 18 krat v 7 delih, na strokovno in znanstveno področje pa skoraj ne zaide. Hrepeneti ima nizko frekvenco, v strokovnem jeziku pa se sploh ne rabi (menda le slučajno ni zajeta v nobeni pesmi[4]). *Nazor* je razen v pesništvu zastopan povsod, toda vidno se dvigne njegova frekvenca v skupini G (pojavi se 130 krat v sedmih delih), kar je odvisno od specifičnosti izpisanega dela. Beseda *negativen* živi

[3] Tu so imeli avtorji pri delu precejšnje težave. Morali so opustiti prvotni načrt, da bi zajeli direktno govorjeni jezik, in se zadovoljiti z zapisanimi govori, ki so izšli v tisku.

[4] V uvodu k slovarju pravi J. Bělič: »Skupno število 1,623.527 besed, s katerimi se je tu delalo, pa še zdaleč ni tako veliko, da ne bi vsaj pri besedah s srednjo frekvenco izključevalo slučajnosti, ki izhaja iz tega, kateri teksti so bili izbrani za ekscerpiranje.« (str. 5). Za besede z nižjo frekvenco lahko to velja tembolj; te se namreč izmikajo zakonitostim in so bolj karakteristične za posamezna stilna področja. V našem primeru bi to veljalo za besedo **hrepeneti.** Možnost slučaja in število k ekscerpciji pritegnjenih tekstov sta si torej v obratnem sorazmerju.

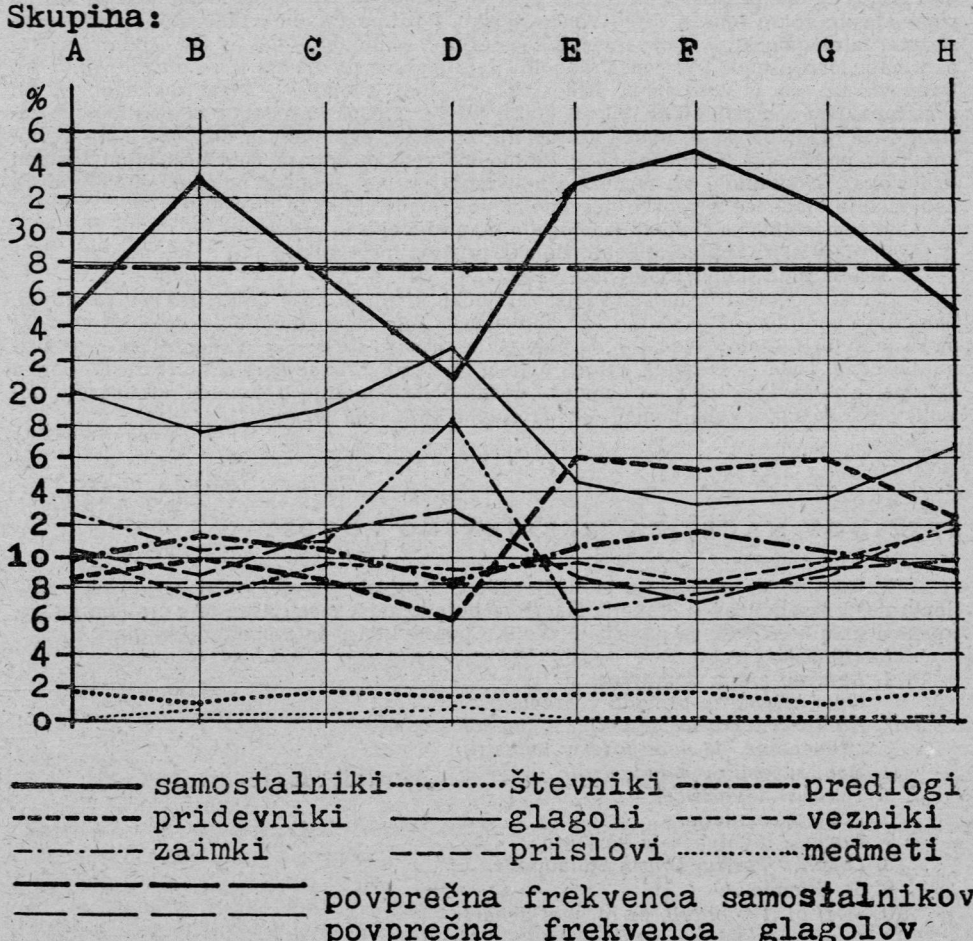
na zelo omejenem stilnem področju jezika, medtem ko se rabi *korelacija* le še v čisto znanstvenih tekstih (v skupini G, kjer se pojavi 71-krat v štirih delih).

Teh številčnih podatkov seveda ne smemo jemati mehanično. S samim upoštevanjem absolutne frekvence številčnega podatka o frekvenci besede ne ovrednotimo pravično; pri tem moramo upoštevati tudi to, v koliko skupinah beseda živi in v koliko virih je izpričana. Visoko absolutno frekvenco lahko kaže beseda zaradi specifičnosti ekscerpiranega dela, medtem ko je lahko beseda z nižjo absolutno frekvenco, vendar izpričana v večjem številu skupin, sestavni del jedra besednega zaklada v jeziku.

Tako razdeli slovar besede glede na frekvenco v pet skupin:

1. splošno rabljene besede z visoko frekvenco (visoka frekvenca, v mnogo virih, v vseh skupinah);
2. splošno rabljene besede z nižjo frekvenco (nižja frekvenca, enakomerno razporejene v razmeroma velikem številu virov, v večini skupin);
3. besede z razmeroma visoko frekvenco, toda le v malo skupinah (npr. têrmini);
4. redke besede z nizko frekvenco, malokrat izpričane.
5. besede z visoko frekvenco, toda izpričane v enem viru, torej tesno vezane na vsebino vira.

S pomočjo številčnih podatkov je mogoče ugotoviti tudi bogastvo in pestrost besednega zaklada pri posameznem delu ali skupini. Tako slovar loči frekvenco gesel in frekvenco besed. Pod geslom razume besedo kot leksikalno enoto; številka, ki označuje



frekvenco gesla, torej kaže, koliko različnih besed je v delu ali skupini uporabljenih. Številka, ki označuje, kolikokrat se je beseda v določenem tekstu ali skupini ponovila, pa kaže frekvenco besede. Razmerje med frekvenco besed, tj. vseh uporabljenih besed, in frekvenco gesel, tj. uporabljenih raznih besed, naziva slovar *indeks ponovitev*, ki do neke mere kaže, kako bogat in pester je besedni zaklad določenega dela ali skupine. Ugotovljeno je, da je indeks ponovitev v skupini B (pesništvo) najmanjši, kar pomeni, da ima pesništvo najpestrejši slovar. Najvišji indeks ponovitev je v skupinah EGH — slovar teh skupin je manj pester (npr. v skupini G), kar izvira iz osredotočenosti na ozek znanstveni problem.

Slovar nadalje analizira in primerja frekvenco besed in frekvenco gesel ter indeks ponovitev pri vseh besednih vrstah.

V posebnem poglavju so na osnovi analize številčnih podatkov zbrani nekateri splošno veljavni zaključki, npr. frekvenčni odnos med samostalniki in glagoli, med pridevniki in samostalniki ter glagoli itd. Izmed številnih diagramov, ki kažejo te odnose, naj pokažemo diagram (gl. str. 147), ki kaže celotno frekvenco posameznih besednih vrst po skupinah (frekvenca besed).

Vidimo, da zgornjo, najmočnejšo plast tvorijo samostalniki in glagoli, v srednji plasti se prepletajo pridevniki, zaimki, prislovi, predlogi in vezniki, spodnjo, neznatno plast pa tvorijo števniki in medmeti. Presenetljivo visoko so zastopani samostalniki v skupini F (žurnalistika), ki so mnogo nad povprečjem frekvence samostalnikov (27,77 %), pa tudi v skupini B (pesništvo) ne bi pričakovali tako močne frekvence samostalnikov, medtem ko je številčnost v skupinah E in G razumljiva zaradi samostalniškega načina izražanja v strokovnem in znanstvenem jeziku. Pozornost vzbudi tudi odnos med samostalniki in glagoli, ki je v obratnem sorazmerju: v skupinah, kjer so samostalniki zastopani najmočneje, pade frekvenca glagolov celo pod povprečje (ki je za glagole 18,15 %). Dalje vidimo, da se zastopanje pridevnikov (četudi v mnogo manjši frekvenci) ujema z zastopanjem samostalnikov (razen v skupini F), isto pa velja tudi za predloge. Z zastopanjem glagolov pa se ujema zastopanje zaimkov, prislovov in do neke mere veznikov (kar preseneča). Glede na medsebojne frekvenčne odnose med besednimi vrstami loči slovar tri skupine, pri čemer je o uvrstitvi v isto skupino odločalo ujemanje ali neujemanje tendence z vodečo besedno vrsto (samostalniki in glagoli):

1. samostalniška skupina (samostalniki, pridevniki in predlogi);
2. glagolska skupina (glagoli, zaimki, prislovi in vezniki);
3. nevtralna skupina (števniki, vezniki).

Tu so navedeni le najsplošnejši zaključki, ki jih prinaša češki frekvenčni slovar. Zanimivih podrobnosti je veliko več. Slovarju bi bilo mogoče očitati pomanjkljivost (ki pa se je avtorji dobro zavedajo), da namreč ne prinaša frekvence pomenov pri posameznih besedah. Toda če bi hoteli avtorji ustreči tej zahtevi, bi se delo silno razraslo, sistem dela pa je zastavljen tako, da omogoča dodatno razlikovanje frekvence posameznih pomenov, pri čemer seveda pomaga dober slovar knjižnega jezika.

Tomo Korošec

PREŽIHOV NAČRT ZA IZSELJENSKO POVEST

V Prežihovi zapuščini se je med različnimi zapiski ohranil načrt za povest o življenju naših izseljencev v Franciji. Načrt vsebuje trinajst točk, pisan je s črnilom na obtrganem listu konceptnega papirja. Nekoliko popravljen v interpunkciji se glasi:

Izseljeniška povest.

1. *Delavec gre iz domovine.*
   Opis odhoda in prihoda v Francijo.
2. *Kontrakt v gozdu, na polju, v kamnolomu.*
3. *Hrepenenje, da dobi posel v industriji.*
4. *Dobi zaposlitev v industriji.*
5. *Tovariši, tovarišice.*
6. *Podjetna policija.*
7. *Domače društvo — s kontrolo.*
8. *Lokali v Parizu, Lensu, Sallaumine.*
9. *St. Etienne.*
10. *»Kaj bi ti napravil, če bi imel denar?«*
11. *Denar poneverjen pri izmenjavi na postaji.*