FILOLOGIJA, KNJIGA 14, ZAGREB 1986. *YU ISSN 0449—363X*

UDK 802.0:970
Izvorni znanstveni članak
Primljeno 2/1987.

ŽELJKO BUJAS

O SREDIŠNJOJ JEZGRI AMERIČKOENGLESKOG VOKABULARA

Jezik je ustrojen na nekoliko uglavnom jasno odvojenih razina (fonetskoj, fonološkoj, morfološkoj, sintaktičkoj, leksičkoj itd.). Svaka od tih razina ima opet svoju strukturu s vlastitim zakonitostima. Tako je i ustrojstvo leksika višestruko: semantičko, stilističko, kronološko, geografsko, frekvencijsko. Predmet ovog rada bit će posljednji od nabrojenih aspekata, čestotni.

Zakonitosti frekvencijske strukture općeg vokabulara tzv. svjetskih jezika dovoljno su uočene zahvaljujući dosad provedenim čestotnim istraživanjima. Neposredno najupotrebljiviji rezultati tih istraživanja su svakako vokabularske rang-liste, to jest popisi leksičkih jedinica popraćenih individualnom i kumulacijskom frekvencijom, od najviše prema nižim vrijednostima. Dosad je objavljeno petnaestak velikih frekvencijskih rang-lista najvažnijih svjetskih jezika: engleskog, ruskog, njemačkog, francuskog, španjolskog, portugalskog. Kako je najtemeljitije istražena frekvencijska struktura američkog engleskog, razmotrit ćemo zakonitosti koje ona očituje.

Analizirat ćemo najnoviju rang-listu američkog engleskog objavljenu u *Frequency Analysis of English Usage* od W. Nelsona Francisa i Henryja Kučere (Houghton Mifflin, Boston, 1982). To je zapravo dorađena ranija lista iz *Computational Analysis of Present-Day American English* istih autora (Brown Univ. Press, Providence, 1967). Zasnovana je dakle na poznatom tzv. Brown Korpusu suvremenog (pisanog) američkog engleskog (od 1 014 000 riječi teksta), čije su temeljne značajke stilistička reprezentativnost, sinkronost i visoka disperzija uzoraka. Stilistička reprezentativnost ostvarena je obuhvatom 15 žanrova: od znanstvenog i filozofskog preko štampe, stručnih tekstova i javnog jezika do beletristike i pop književnosti. Sinkronost je osigurana time što su uključeni samo tekstovi objavljeni (prvi put) u 1961. Disperzija uzoraka — da se izbjegne tematska koncentracija leksičkih jedinica — postignuta je velikim brojem uzoraka (ukupno 500, svaki s oko 2 000 riječi /6 tipkanih stranica/ teksta).

Čestotna lista Brown Korpusa suvremenog američkog engleskog iz 1982. predstavlja dorađenu verziju ranije liste iz 1967, utoliko što je ra-

nija lista sada »lematizirana«, to jest razriješeni su homografi, a kosi oblici podvedeni su pod temeljni, kanonski oblik. Lematizacija koju su jednostavni kompjuterski postupci iz 1967. ignorirali, provedena je sada potpuno i uglavnom kompjuterski (u 22% slučajeva još uvijek je naime izvršena ručno). Lematizirana lista (1982) svakako je upotrebljivija za većinu leksičko-vokabularskih analiza i primjena, s time da se podaci o frekvenciji kosih oblika s nezavisnom gramatičkom funkcijom mogu i dalje pronaći u abecednoj listi leksika korpusa (koja zauzimlje četiri petine izdanja iz 1982).

Pokušajmo sada uočiti zakonitosti frekvencijske strukture američko-engleskog vokabulara na osnovi (lematizirane) čestotne liste Brown Korpusa. Razmotrit ćemo podrobno samo prvih sto riječi s liste, na što smo upućeni prostornim ograničenjima ali i opravdanim lingvističkim razlozima koji će se ubrzo pokazati:

Rang (redni broj)	Leksička jedinica	Apsolutna frekvencija (ukupni broj javljanja u Korpusu)	Relativna frekvencija (% od ukupnog broja riječi u Korpusu)	Kumulativna (pribrajana) frekvencija
1	2	3	4	5
1	the	69 975	6,903	—
2	be	39 175	3,865	10,768
3	of	36 432	3,594	14,362
4	and	28 872	2,848	17,210
5	a	23 073	2,276	19,846
6	in	20 870	2,059	21,545
7	he	19 427	1,917	23,462
8	to (inf)	15 025	1,482	24,944
9	have	12 458	1,229	26,173
10	to (prep)	11 165	1,110	27,283
11	it	10 942	1,079	28,362
12	for	8 996	0,885	29,247
13	I	8 387	0,827	30,074
14	they	8 284	0,817	30,891
15	with	7 286	0,719	31,610
16	not	6 976	0,688	32,298
17	that (conj)	6 468	0,638	32,936
18	on	6 183	0,610	33,546
19	she	6 039	0,596	34,142
20	as	6 029	0,595	34,731
21	at	5 377	0,530	35,267
22	by	5 246	0,518	35,785
23	this	5 145	0,508	36,293
24	we	4 865	0,480	36,773
25	you	4 620	0,456	37,229
26	from	4 371	0,431	37,660

1	2	3	4	5
27	do	4 367	0,431	38,091
28	but	4 226	0,417	38,508
29	or	4 204	0,415	38,923
30	an	3 727	0,368	39,291
31	which	3 560	0,351	39,642
32	would	3 062	0,302	39,944
33	say	2 765	0,273	40,217
34	all	2 758	0,272	40,489
35	one	2 737	0,270	40,759
36	will	2 686	0,265	41,024
37	who	2 678	0,264	41,288
38	that (dem)	2 455	0,242	41,530
39	when	2 333	0,230	41,760
40	make	2 312	0,228	41,988
41	there (exist)	2 280	0,225	42,213
42	if	2 199	0,217	42,430
43	can (aux)	2 192	0,216	42,646
44	man	2 110	0,208	42,854
45	what	1 955	0,193	43,047
46	time	1 901	0,188	43,235
47	go	1 844	0,182	43,417
48	no	1 821	0,180	43,597
49	into	1 790	0,177	43,774
50	could	1 782	0,176	43,950
51	up	1 712	0,169	44,119
52	other	1 710	0,169	44,288
53	that (pron)	1 674	0,165	44,453
54	year	1 661	0,164	44,617
55	out	1 644	0,162	44,779
56	new	1 635	0,161	44,940
57	some	1 594	0,157	45,097
58	take	1 575	0,155	45,252
59	these	1 575	0,155	45,407
60	come	1 561	0,154	45,561
61	see	1 513	0,149	45,710
62	get	1 486	0,147	45,857
63	know	1 473	0,145	46,002
64	state (sub)	1 421	0,140	46,142
65	two	1 412	0,139	46,281
66	only	1 360	0,134	46,415
67	then	1 348	0,133	46,548
68	any	1 335	0,132	46,680
69	now	1 314	0,130	46,810
70	may	1 307	0,129	46,939

1	2	3	4	5
71	than	1 297	0,128	47,067
72	give	1 264	0,125	47,192
73	about	1 242	0,123	47,315
74	as (qual)	1 101	0,109	47,424
75	day	1 077	0,106	47,530
76	also	1 070	0,106	47,636
77	find	1 033	0,102	47,738
78	first	1 031	0,102	47,840
79	way	1 027	0,101	47,941
80	must	1 017	0,100	48,041
81	use (vb)	1 016	0,100	48,141
82	more (qual)	1 015	0,100	48,241
83	like (conj)	1 012	0,100	48,341
84	even (adv)	997	0,098	48,439
85	many	997	0,098	48,537
86	more (postdet)	990	0,098	48,635
87	think	982	0,097	48,732
88	such	978	0,097	48,829
89	where	949	0,094	48,923
90	so	932	0,092	49,015
91	through	926	0,091	49,106
92	should	915	0,090	49,196
93	people	902	0,089	49,285
94	each	878	0,087	49,372
95	those	864	0,085	49,457
96	M(iste)r	857	0,085	49,542
97	over	843	0,083	49,625
98	world	832	0,082	49,707
99	seem	831	0,082	49,789
100	just	795	0,078	49,867

Prvi mogući zaključak je da ovih 100 najčešćih riječi »pokriva« polovicu, točno 49,4%, svakog (prosječnog) teksta na američkom engleskom. Prema tome, zamišljajući ustrojstvo vokabulara u koncentričnim krugovima, zaista možemo nazvati ovih 100 riječi njegovom središnjom jezgrom *(core vocabulary)*. To je međutim statičko zapažanje, a mnogo je značajniji dinamizam frekvencijske strukture što ga uočavamo prikazujući listu u segmentima po 10 jedinica (uz pojednostavljene vrijednosti u %; v. slijedeću tabelu).

Očit je drastičan pad kumulativnosti između prve i druge desetine s liste, za čime slijedi kontinuirani ali sve blaži pad kumulativne stope. Dok prvih 50 riječi ukupno pokriva 44,0% Korpusa, drugih 50 donosi samo 5,9%. Taj trend postaje još jasniji kad prikažemo »pokrivenost« Korpusa i kumulativne stope za dalje stotine i tisuće na rang-listi Brown Korpusa. Ovdje ćemo se doduše poslužiti podacima s nelematizirane

Rang	Ukupna apsolutna frekvencija	Ukupni % pokrivenosti Korpusa	Kumulativni %
1— 10	276 472	27,3	—
11— 20	75 590	7,4	34,7
21— 30	46 148	4,6	39,3
31— 40	27 346	2,7	42,0
41— 50	19 874	2,0	44,0
51— 60	16 341	1,6	45,6
61— 70	13 969	1,3	46,9
71— 80	11 159	1,1	48,0
81— 90	9 868	1,0	49,0
91—100	8 643	0,9	49,9

frekvencijske rang-liste Brown Korpusa iz 1967, budući da ta lista sadrži precizno (kompjuterski) izračunate kumulativne stope za svih 50 406 različitih jedinica na 1 013 644 riječi korpusnog teksta. Lematizirana lista iz 1982. ne donosi kumulativne stope, što je ozbiljan metodološki nedostatak i znatno otežava čestotne analize te liste. Kumulativne stope za prvih 100 jedinica s liste morao sam stoga ručno izračunati. Ručno izračunavanje za kasnije stotine i tisuće moralo se, očito, isključiti. Razlike kumulativnih vrijednosti između dvije liste postoje: na primjer 47,4% za prvih 100 jedinica nelematizirane liste prema naših 49,9%. One međutim nisu značajne, pogotovo na nižim dijelovima rang-liste gdje se kumulativne stope svode na sve beznačajnije vrijednosti.

TABELA KUMULATIVNOSTI ZA PRVIH 1000 JEDINICA (PO 100)

Rang	% pokrivenosti Korpusa	Kumulativna stopa (+ %)
1— 100	47,43	—
— 200	53,59	6,16
— 300	57,21	3,62
— 400	59,86	2,65
— 500	61,93	2,07
— 600	63,68	1,75
— 700	65,19	1,51
— 800	66,49	1,30
— 900	67,61	1,22
—1000	68,83	1,12

TABELA KUMULATIVNOSTI ZA PRVIH 10 000 JEDINICA (PO (1000)

Rang	% pokrivenosti Korpusa	Kumulativna stopa (+ %)
1— 1000	68,8	—
— 2000	76,2	7,4
— 3000	80,7	4,5
— 4000	83,7	3,0
— 5000	85,8	2,1
— 6000	87,5	1,7
— 7000	88,9	1,4
— 8000	90,1	1,2
— 9000	91,0	0,9
—10000	91,8	0,8

TABELA KUMULATIVNOSTI ZA SVIH 50 406 JEDINICA (PO 10 000)

Rang	% pokrivenosti Korpusa	Kumulativna stopa (+ %)
1—10 000	91,8	—
10 001—20 000	96,2	4,4
20 001—30 000	98,0	1,8
30 001—40 000	99,0	1,0
40 001—50 406	100,0	1,0

Kretanje, to jest opadanje, stopa kumulativnosti nesumnjivo je temeljna zakonitost frekvencijske strukture vokabulara američkog engleskog, od njegove središnje jezgre pa dalje prema vanjskim koncentričnim slojevima, prema jedinicama sve niže učestalosti. Ista zakonitost vrijedi dakako i za vokabulare drugih jezika.

Ovo svojstvo vokabularske strukture ima svoje praktične pedagoške vrijednosti, pa i psihološke implikacije. Učeći strani jezik godinu dana, mi naime svladavamo oko tisuću riječi — u pravilu najčešćih. S tako usvojenim vokabularom pokrivamo (praktički: razumijemo) blizu 70% generalnog teksta na tom jeziku, u našem slučaju na američkom engleskom. Međutim, nastavljajući učenje još jednu godinu, dakle udvostručujući napor, mi ćemo, savladavši tako 2 000 riječi, pokrivati tek nešto više od 75% prosječnog američkoengleskog teksta. Time smo dakle povećali svoju sposobnost razumijevanja takvog teksta za svega 7,4%. Ova vokabularska svojstva, sa svojim jasno ugrađenim ograničenjima, bolje od bilo čega objašnjavaju zašto prosječni tečaj stranog jezika ne traje više od dvije godine. Nakon toga, omjer uloženog vremena (napora) i postignutih rezultata sve je nepovoljniji; učenje jezika u obliku jezičnog tečaja postaje neekonomično. Sada se naime ulazi u fazu učenja kad se svladavanje samog vokabulara postavlja kao imperativ, a konzultiranje rječ-

nika i čitanje tekstova na stranom jeziku postaje temeljnim postupkom. Naravno, pokrivanje i razumijevanje teksta nije sasvim isto, pa tako osoba s pasivnim vokabularom od 10 000 (najčešćih) riječi američkog engleskog razumije, zahvaljujući kontekstu, više od 91,8% teksta.

Zanimljive su i neke druge implikacije čestotne strukture središnje jezgre američkog engleskog.

Tako se ta jezgra, što je opće svojstvo vokabulara, sastoji pretežno od »funkcionalnih« ili »gramatičkih« jedinica. Leksičkih jedinica (»punih« riječi), to jest imenica, glagola, pridjeva i pridjevskih priloga, ima ukupno tek 25, a to su:

imenice: *man, time, year, state, day, way, people, (Mister)* i *world*

glagoli: *(be), (have), (do), say, make, go, take, come, see, get, know, give, find, use, think* i *seem*

pridjevi: *new*

Ako imenice najneposrednije izražavaju pojmovni inventar jezika, može se ustvrditi da najčešće imenice izražavaju temeljne pojmove i ukazuju na osnovne čovjekove preokupacije. To su, prema našoj središnjoj jezgri, on sam *(man, people)* u vremenu *(time, year, day)* i prostoru *(way, world)*. Univerzalnost ovog zapažanja potvrđuje i središnja jezgra ruskog vokabulara. Tamo su, u vrlo sličnom korpusu (L. N. Zasorina, *Častotnyj slovar' russkogo jazyka*, 1977), najčešće imenice: *god, delo, vremja, čelovek, ljudi, ruka, žizn', den', tovarišč, rabota* i *glaz*. Zanimljivo je da su *tovarišč* i njegov kulturni ekvivalent u američkom engleskom *Mister* na gotovo identičnom mjestu rang-liste (92-om odnosno 94-om).

Konačno, ne manje važno svojstvo središnje jezgre američkoengleskog vokabulara, koje se podjednako nameće, jest i njegov etimološki sastav. Ništa jasnije ne posvjedočuje anglosasku, germansku bit engleskog jezika (naravno i u njegovoj daleko najbrojnijoj američkoj varijanti) od činjenice da od njegovih najčešćih 100 riječi samo 8 nisu anglosaske. Od toga su 3 staroskandinavske (*they, take, get*; ukupno 1,119%) a 5 starofrancuske odnosno anglofrancuske *(state, use, Mister, just, people)*. Ukupni udio ovih drugih je manji od polovice jednog postotka (precizno: 0,492%), što sve čini središnju jezgru američkog engleskog preko 99,5-postotno germanskom ili, uže gledano, 98,4-postotno anglosaskom. Već u drugoj stotini rang-liste, kao što je poznato iz dosadašnjih istraživanja,[1] počinje opadati udio anglosaskog elementa u vokabularu. On pada ispod 50% u petoj stotini i spušta se na 33% u desetoj stotini. Unatoč tome, utjecaj mase središnje jezgre je toliko značajan da je u ukupnom etimološkom sastavu prve tisuće riječi u američkom engleskom anglosaski udio još uvijek čak 85,1%, francuski tek 8,4%, latinski 2,6%, staroskandinavski 2,4% i grčki 1,2%.

[1] A. Hood Roberts, *A Statistical Linguistic Analysis of American English*, Mouton, 1965.

Željko Bujas, »An Analysis of the Etymological Makeup of the English Core Vocabulary«, *Studia Romanica et Anglica Zagrabiensia*, 41—42/1976, 25—42.

Bilo bi posebno zanimljivo usporediti središnje jezgre američkog i britanskog engleskog, što je sada moguće zahvaljujući najnovijim frekvencijskim analizama sumjerljivog Korpusa britanskog engleskog.[2] To bi međutim zahtijevalo uspoređivanje rang-lista daleko preko raspona 100 najčešćih riječi pa traži poseban rad.

Summary

ANALYZING THE CORE VOCABULARY OF AMERICAN ENGLISH

This paper sets out to analyze the frequency structure of the »core vocabulary« of American English. The material analyzed is the lematized version (1982) of the Brown Corpus. Its core vocabulary is defined as the first 100 words on the Rank List, covering 49.9% of any average text in contemporary written American English. The principal observation of this analysis was the steady decline of Rank List cumulative frequencies (and their implication for foreign language tuition and acquisition). A speculative link was also established between the most frequent nouns in American English (and Russian) and fundamental human concerns: Man in time and space. Finally, an almost exclusively Anglo-Saxon and Germanic character has been established for the core vocabulary analyzed, and defined as a significant property of American English.

[2] Knut Hofland i Stig Johansson, *Word Frequencies in British and American English*, The Norwegian Computing Centre for the Humanities, Bergen, 1982.