UDC 801.321:301.008
Original scientific paper
Accepted for publication on 26 September 1988

Cultural-Anthropological Relevance of Word Frequency Parameters (American vs. British English)

Željko Bujas
Faculty of Philosophy, Zagreb

Reliable word counts, now available for major European languages, reveal frequential patterning as a significant constituent of their lexical structure. This patterning shows an intriguing uniformity and has a striking cultural-anthropological relevance. The ten or so nouns, topping the rank lists of these languages consistently refer to the concepts belonging to what this paper terms "Man's fundamental preoccupations" (Man, Time, Space). American-English and British- English rank lists show, as expected, a close correspondence of both the nouns occurring and their rank positions. As one goes down the rank lists (into the second and third hundred), the rank position divergence between nouns found on the lists is encountered. A pronounced or drastic divergence is a sure indication of cultural differences (as illustrated in the paper), and as such an efficient tool of research into the lexical differences between American English and British English.

1.

1.1. Language is structured on several largely separate levels: phonetic, phonological, morphological, syntactic, lexical et cetera. In their turn, each of these levels has its own structure governed by own laws. Consequently, lexis can be viewed as internally organized – semantically, stylistically, chronologically, geographically, frequentially. This paper proposes to deal with the last of these aspects: frequency as a law of lexical structure, specifically exploring its cultural-anthropological implications.

1.2. Laws governing the frequency structure of what are usually termed world languages have been adequately determined for their current general vocabularies. Thanks to a number of large word counts carried out so far (some 15), we now have rank lists of such vocabularies of English, Russian, German, French, Spanish, Portuguese. English, specifically American English, is the best researched of them, with Kučera and

Francis' analyses of the so-called Brown Corpus (in 1967 and 1982), setting the standard for all subsequent large word counts.[1]

1.3. The Brown Corpus Rank List is the product of computer-processing a one-million-word corpus, aiming to be as representative as possible of contemporary written American English. This Rank List is a unique statement of the frequency structure of the lexis of American English. From a purely quantitative point of view, it reveals some striking patterns. So, 100 most frequent words account for close to one-half (47.4%) of any average American English text. This, of course, does not mean that you can *understand* 50% of such a text by knowing merely the 100 most frequent words. But as we go down the list, the degree of text coverage becomes increasingly (and before long practically) equal with the degree of understanding it. This has some very practical implications, as the following table immediately reveals:

BROWN CORPUS

a) **Cumulative Frequency Table** for 10,000 first items (by 1,000)

Rank		Text Coverage (in %)	Coverage Increase (in %)
1	– 1,000	68.8	–
	– 2,000	76.2	7.4
	– 3,000	80.7	4.5
	– 4,000	83.7	3.0
	– 5,000	85.8	2.1
	– 6,000	87.5	1.7
	– 7,000	88.9	1.4
	– 8,000	90.1	1.2
	– 9,000	91.0	0.9
	–10,000	91.8	0.8

b) **Cumulative Frequency Table** for all 50, 406 items (by 10,000)

1	– 10,000	91.8	–
	– 20,000	96.2	4.4
	– 30,000	98.8	1.8
	– 40,000	99.0	1.0
	– 50,406	100.0	1.0

1. W. Nelson Francis and Henry Kučera, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, R.I., 1967 and *Frequency Analysis of English Usage*, Houghton Mifflin, Boston, Mass., 1982.

The above table is best called the Table of Diminishing Returns (to borrow a term from economics), as it makes devastatingly clear why most efforts to learn English stop between the second and third thousand word levels, or else why the typical course of English does not go beyond its second year.

2.

2.1. More relevant, however, from a cultural-anthropological point of view is the Brown Corpus Rank List (from 1967) as a source of frequency structure information about *lexical* items. Let us take a close look at this (simplified) presentation of the first 100 items from this list:

the	for	this	which	we
of	it	had	one	him
and	with	not	you	been
to	as	are	were	has
a	his	but	her	when
in	on	from	all	who
that	be	or	she	will
is	at	have	there	more
was	by	an	would	no
he	I	they	their	if
10	20	30	40	50

out	can	then	man	must
so	only	do	me	* * F
said	other	first	even	through
what	new	any	most	back
up	some	my	made	years
its	could	now	after	where
about	time	such	also	much
into	these	like	did	your
than	two	our	many	way
them	may	over	before	well
60	70	80	90	100

This list makes it obvious that most of its items are functional ("grammatical") words. Only a few are fully or partly lexical in character, with nouns among them (4) picked out in bold print. They are, in order of frequency: *time, man, years* and *way*.

2.2. A "lemmatized" version of the Brown Corpus Rank List (from 1982) is a slightly reordered list – with all oblique forms grouped under the basic entry and with homography taken care of:

the	it	at	which	there *exist*
be	for	by	would	if
of	I	this	say	can *aux*
and	they	we	all	**man**
a	with	you	one	what
in	not	from	will	**time**
he	that *conj*	do	who	go
to *inf*	on	but	that *dem*	no
have	she	or	when	into
to *prep*	as	an	make	could
10	20	30	40	50

up	see	than	use *vb*	through
other	get	give	more *qual*	should
that *pron*	know	about	like *conj*	**people**
year	**state** *sub*	as *qual*	even *adv*	each
out	two	**day**	many	those
new	only	also	more *postdet*	**Mister**
some	then	find	think	over
take	any	first	such	**world**
these	now	**way**	where	seem
come	may	must	so	just
60	70	80	90	100

As we can see, most lexical items are now higher on the list and, as a consequence, we now encounter nine nouns among the first 100 items: *man, time, year, state, day, way, people, Mister* and *world*.

3.

3.1. Now if nouns most immediately express the conceptual inventory of language, it can be claimed that **the most frequent nouns indicate Man's fundamental preoccupations.** These preoccupations, according to both versions of the Brown Corpus Rank List, are Man himself (*man, people*) in Time (*time, year, day*) and Space (*way, world*). The universality of this observation is corroborated by the top frequency nouns in a comparable Russian word count[2]: *god* (year), *delo* (affair, matter), *vremya* (time), *chelovek* (man, person), *lyudi* (people), *ruka* (hand), *zhyzn'* (life), *den'* (day), *tovarishch* (comrade, Mister), *rabota* (work) and *glaz* (eye). *Tovarishch* and its American cultural equivalent *Mister* are uncannily close on their respective rank lists: in 92nd and 94th positions (93rd on the lemmatized rank list, after 10 occurrences of *Mr.* have been run together with *Mister*).

3.2. An intriguing way to test this claim is to compare the ordering of lexical items (specifically nouns) on the rank lists of American and British English, that is of two linguistically identical but culturally different entities. This comparison is now possible thanks to the so- called LOB (Lancaster – Oslo/Bergen) Corpus of British English, published in 1982 and closely modeled on the Brown Corpus.[3]

4.

4.1. Consulting the LOB Rank List (unlemmatized!), we quickly establish the practically identical muster of nouns/concepts for the top 100 items: *time, Mr., man, years, people* and, cheating by only one rank position, *way* (101st on the LOB list).

4.2. Intriguing insights are provided by a comparison of the rank positions of American English items. (Brown Corpus 1967 Rank List) and those in British English (LOB Rank List) in the following table:

	AE	BE	AE>BE		AE	BE	AE>BE
the	1	1	–	out	51	53	+2
of	2	2	–	so	52	46	–6
and	3	3	–	said	53	52	–1
to	4	4	–	what	54	55	+1
a	5	5	–	up	55	56	+1
in	6	6	–	its	56	70	+14
that	7	7	–	about	57	54	–3
is	8	8	–	into	58	62	+4
was	9	9	–	than	59	64	+5
he	10	12	+2	them	60	60	–

2. L.N. Zasorina (ed.), *Chastotnyj slovar' russkogo jazyka*, Russkij jazyk, Moscow, 1977.

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

for	11	11	–	can	61	61	–
it	12	10	–2	only	62	58	–4
with	13	14	+1	other	63	69	+6
as	14	13	–1	new	64	84	+20
his	15	18	+3	some	65	57	–8
on	16	16	–	could	66	65	–1
be	17	15	–2	time	67	63	–4
at	18	19	+1	these	68	71	+3
by	19	20	+1	two	70	74	+4
I	20	17	–3	may	70	74	+4
this	21	22	+1	then	71	68	–3
had	22	21	–1	do	72	73	+1
not	23	23	–	first	73	78	+5
are	24	27	+3	any	74	75	+1
but	25	24	–1	my	75	59	–16
from	26	25	–1	now	76	72	–4
or	27	31	+4	such	77	86	+9
have	28	26	–2	like	78	83	+5
an	29	34	+5	our	79	82	+3
they	30	33	+3	over	80	80	–
which	31	28	–3	man	81	88	+7
one	32	38	+6	me	82	66	–16
you	33	32	–1	even	83	96	+13
were	34	35	–1	most	84	92	+8
her	35	29	–6	made	85	77	–8
all	36	39	+3	after	86	87	+1
she	37	30	–7	also	87	97	+10
there	38	36	–2	did	88	107	+19
would	39	43	+4	many	89	94	+5
their	40	41	+1	before	90	91	+1
we	41	40	–1	must	91	85	–6
him	42	49	+7	"any formula"	92		
been	43	37	–6	through	93	120	+27
has	44	42	–2	back	94	102	+8

when	45	44	−1	years	95	90	−5
who	46	50	+4	where	96	93	−3
will	47	48	+1	much	97	89	−8
more	48	51	+3	your	98	111	+13
no	49	47	−2	way	99	101	+2
if	50	45	−5	well	100	95	−5

The rank position agreement is remarkable. As could be expected, it is total for the first 9 items on the list, but also for another 6 items scattered down the lists, with one item (*over*) as far down as the 80th position. Of the remaining 85 items, as many as 77 stay within 10 positions from each other. Of the 8 items diverging in excess of 10 positions, the three most divergent – *through* (+27), *new* (+20) and *did* (+19) – show consistent higher positions in AE. Though not drastic, this divergence tempts one to offer explanations for the higher frequency of the AE item. The higher rank od *did*, for instance, is probably due to the more frequent use of the past tense in AE. Indeed, *have* and *has* both show a slightly lower (−2) rank position.

4.3. A comparison of nouns on the two lists also shows a remarkable rank position agreement:

	AE	BE	AE>BE
time	67	63	−4
Mr.	108	76	−32
man	81	88	+7
years	95	90	−5
way	99	101	+2
people	107	100	−7

The sole exception of *Mr.*, considerably higher in British English, is intriguing, hinting at cultural differences but more about this later on.

4.4. With the second hundred of items, the two rank lists are still remarkably close, sharing nouns expressing the same fundamental preoccupations of Man as those from the first 100 rank positions (remember: Man himself, Man in Time and Man in Space). These nouns are: *world, men, life, day, year, house* and *home*. A few other nouns expand the fundamental concept areas: *part, school, work, number, course* and *fact*. Another few items go beyond the 200th rank position on the British list – either negligibly so (*government, hand*) or staying below the 250th position (*use, school, war, water*).

Rank position differences between all these items on the two lists are less than 50 rank steps. For six of them (*people, life, year, house, number, hand*) they are remarkably close (10 steps or less.) Only two items (*use* and *place*) come near diverging by 100 rank positions. Consequently, any drastic rank position divergence between the two lists must be an indication of other forces at play beyond mere lexical ordering. With the American and British rank lists largely comparable in frequency structure, any striking structural

dislocation must be seen as a signal of disparate cultural content of the lexical items involved. A pattern of similarly diverging pairs is, in its turn, evidence of an underlying "Conceptual Rank List" out of synch with the 'surface' lexical rank list.

In our case – with the second hundred American and British rank list items (nouns) being compared – only two items showed such drastic rank divergence. They were: *state*, with its AE rank (112th) topping the BE rank by 346 positions (the plural form *states* revealed an even more extreme span of 836 positions), and the BE *sir* whose rank of 195th was six times higher than its position in AE (1125th). The cultural mechanisms behind the much higher ranking of the item/concept *state(s)* on the AE list, or of *sir* on the BE list, are fairly obvious.

4.5. The third hundred of rank list items reveals, as expected, a growing divergence between the two lists. Still, a total of 19 items stay within the 50 rank span (*church*, with a shift of only 6 steps, being the concept of closest correspondence). The extremest cases of divergence, in favor of the AE item, were: *program* (+659 rank positions), *city* (+275), *president* (+224), *development* (+199) and *business* (+196). While *president* and, probably, *city* could be explained as civilizational rank shifts, one is tempted to interpret *program, development* and *business* as 'buzz words' specific for AE. Similarly diverging cases, in favor of BE item, are harder to explain. *Council*, diverging most extremely (+766), can probably be put down to civilization, but *view* (+221) and *book* (+204) could hardly be explained away as British buzz words. One is also intrigued by the significantly higher (+126) position of *girl* in BE, in view of the consistently lower position of female pronouns *she, her, hers* on both lists compared to *he, his, him*.

5.

An entirely different direction of analysis is possible with synonym pairs. These may involve synonyms proper (such as *automobile* 2 /BE/ : 100 /AE/, but *car* 335 : 393), or cultural pairs (*coffee* 54 : 78 and *tea* 111 : 28; or *ocean* 12 : 34 and *sea* 159 : 95). Important stylistic insights can be gained by analyzing the frequency skew in such pairs as *perhaps* (406 /BE/ : 307 /AE/) and *maybe* (82 : 134). But single word skews are equally revealing: *trousers* (25 : 9), *mighty* (17 : 29), *guess* (35 : 56), and the like.

6.

More sophisticated analytical techniques than a mere comparison of total absolute frequencies are possible even with the LOB and Brown Corpus. Rank correlation between the 15 text categories is an obvious next procedure, and the LOB compilers have taken the first steps there. Still, one has to bear in mind all the time that a one-million-word corpus is simply too small to be completely reliable. After considerably larger corpora are made available in the future (with desktop optical readers and personal supercomputers in many a linguist's den), we will be able to look deeper into this intriguing question of full cultural implications of lexical frequency structure.

KULTURNO-ANTROPOLOŠKA RELEVANTNOST FREKVENCIJSKIH PARAMETARA RIJEČI (AMERIČKI ENGLESKI NASUPROT BRITANSKOM ENGLESKOM)

Pouzdani čestotni rječnici, kakvi sada postoje za najvažnije evropske jezike, pokazuju frekvencijske obrasce koji se značajno uklapaju u leksičko ustrojstvo tih jezika. Takvi obrasci privlače pažnju svojom uniformnošću koja se nameće kao kulturno- antropološki relevantna. Desetak imenica s vrha ranglista spomenutih jezika dosljedno izražavaju pojmove koji pripadaju, kako to ovaj rad nazivlje, »temeljnim čovjekovim preokupacijama« (Čovjek, Vrijeme, Prostor). Rangliste američkog engleskog i britanskog engleskog, kao što se moglo očekivati, tijesno se poklapaju – kako u odabiru imenica tako i u njihovom mjestu na ranglisti. Nešto niže na ovim listama, u drugoj i trećoj stotini riječi, susrećemo raskorak u položaju (rangu) prisutnih imenica. Izraziti odnosno drastični raskoraci sa sigurnošću ukazuju na kulturne razlike (što se ilustrira primjerima u radu) i tako predstavljaju djelotvoran istraživački postupak za analize američko-britanskih leksičkih razlika .