LETTER AND WORD FREQUENCIES

REX GOOCH
Letchworth Garden City, Herts, England
rexgooch@ntlworld.com

Counts of letters and words have been made from various bodies of text over the years. Perhaps the earliest well-known corpus was that assembled by W.N. Francis and H. Kucera of Brown University (hence known as the Brown corpus). It contains one million words of American English from 1961 in 15 categories of 500 texts. If it has a use today, it is because matching studies were done of British English (LOB corpus) and Indian English (Kolhapur Corpus) among others. In *Word Ways*, I have preferred to use the Collins COBUILD corpus, known as the Bank of English®: it runs to hundreds of millions of words of English text from British, US, Australian, and Canadian sources (including textbooks, novels, newspapers, guides, magazines, and websites). The corpus has been automatically word-class tagged, and a 200-million-word corpus has been parsed. It is at http://www.cobuild.collins.co.uk/. Recently, I have used the British National Corpus, which is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written (available from http://www.natcorp.ox.ac.uk/, the Oxford University website). There is a Baby version, and the original Brown corpus, tagged, is also available.

In the February issue of *Word Ways* (page 41), Anil remarked that the order of popularity of letters varied greatly depending on whether the letters are counted from running text, or from a list of the most popular words. I believe he was talking about column 6 below, to which I have added actual frequencies. He was mainly comparing the order with that in column 4, the Linotype machine. The Linotype order is very close to that of the BNC running text in column 3. There is a significant difference in *order* between letters and running text. However, in all cases E is dominant, followed by the same cluster of letters — differences in frequency are small in this cluster, so letters easily change places, and the differences in *frequency* are not truly significant. I thought that perhaps any difference betwen running text and words might be due to a difference between the most popular words (often repeated in running text), and others. I therefore decided to count the letters in the most popular 999 words in the BNC, then in the next 85,801: the results in columns 1 and 2 do in my view show significant differences, eg in the more popular words (which are shorter), E is more frequent, and H less so. Similarly, T seems to be more common in running text. I leave the reader to browse the results further. Column 5 is from the same source as Anil's column 6, but has no plural words or words with common suffixes: the lack of terminal -S and -ING are just apparent.

Finally, I include the top 100 words from the BNC, and from the American Heritage Word Frequency Book (Carroll, Davies and Richman, 1971). It draws upon 5 million running words used in US schools. The numbers are the number of occurrences per 10,000 words for the BNC.

I trust the tables will also serve as a reference for *Word Ways* contributors.

Parts per thousand for individual letters in various selections of text

1		2		3		4	5		6	
E	134	E	108	E	125	E	E	114	E	117
A	73	A	85	T	92	T	A	85	I	86
T	73	I	83	A	81	A	I	79	S	81
R	73	S	76	O	76	O	R	75	A	79
O	71	R	73	I	73	I	T	74	R	74
N	70	N	72	N	70	N	O	71	N	74
S	67	T	65	S	63	S	N	64	T	67
I	65	O	64	R	61	H	S	55	O	59
L	50	L	60	H	54	R	L	55	L	53
D	41	C	41	L	41	D	C	47	C	41
C	39	D	38	D	39	L	U	36	D	38
U	33	U	32	C	31	U	P	32	U	32
H	32	M	30	U	28	C	M	32	G	28
M	31	G	28	M	24	M	D	31	P	27
P	28	P	27	F	22	F	H	27	M	27
G	24	H	26	P	20	G	G	23	H	22
Y	21	B	20	G	20	Y	B	21	B	20
W	18	Y	18	W	20	P	Y	20	Y	17
F	17	F	14	Y	19	W	F	14	F	14
B	14	K	11	B	16	B	V	10	V	11
V	12	V	11	V	10	V	W	9	K	9
K	9	W	10	K	7	K	K	8	W	9
J	1	Z	4	X	2	X	X	3	Z	5
X	1	X	3	J	2	J	Z	2	X	3
Q	1	J	2	Q	1	Q	Q	2	J	2
Z	0	Q	2	Z	1	Z	J	1	Q	2

1. top 999 words in BNC (5.5 letters per word)
2. next 85801 words in BNC (7.7 letters per word)
3. running text, BNC
4. Linotype
5. 18584 Common Base words
6. 45406 Common Words

BNC		AmHer
THE	651	the
OF	309	of
AND	282	and
TO	269	a
A	226	to
IN	198	in
THAT	117	is
IT	114	you
IS	105	that
WAS	97	it
I	95	he
FOR	89	for
ON	76	was
YOU	73	on
HE	71	are
BE	69	as
WITH	68	with
AS	54	his
BY	53	they
AT	50	at
HAVE	49	be
ARE	49	this
THIS	48	from
NOT	48	I
BUT	47	have
HAD	46	or
HIS	45	by
THEY	45	one
FROM	43	had
SHE	40	not
WHICH	39	but
OR	39	what
WE	37	all
AN	36	were
THERE	34	when
HER	34	we
WERE	33	there
ONE	30	can
DO	29	an
BEEN	28	your
ALL	27	which
THEIR	27	their
HAS	27	said
WOULD	26	if
WILL	26	do
WHAT	26	will
IF	24	each
CAN	24	about
WHEN	22	how
SO	22	up
NO	22	out
SAID	22	them
WHO	21	then
MORE	21	she
ABOUT	20	many
UP	19	some
THEM	18	so
SOME	18	these
COULD	17	would
HIM	17	other
INTO	17	into
ITS	17	has
THEN	16	more
TWO	16	her
OUT	16	two
TIME	16	like
LIKE	16	him
ONLY	16	see
MY	16	time
DID	15	could
OTHER	14	no
ME	14	make
YOUR	14	than
NOW	14	first
OVER	13	been
JUST	13	its
MAY	13	who
THESE	13	now
NEW	13	people
ALSO	13	my
PEOPLE	13	made
ANY	13	over
KNOW	12	did
VERY	12	down
SEE	12	only
FIRST	12	way
WELL	12	find
AFTER	12	use
SHOULD	11	may
THAN	11	water
WHERE	11	long
BACK	10	little
HOW	10	very
GET	10	after
MOST	10	words
WAY	10	called
DOWN	10	just
OUR	9	where
MADE	9	most
GOT	9	know