LETTER FREQUENCIES AND WORD LENGTHS

REX GOOCH
Welwyn, Herts, England

LETTER FREQUENCIES IN DICTIONARIES AND RUNNING TEXT

In trying to find an explanation for a certain phenomenon, I decided to compare the frequencies of letters in a certain group of words with some norm. I was not too happy with available lists because of their small sample size, and also because it was difficult to see whether differences in rank between different lists were significant. Therefore I made my own. The table in this article is compiled from approximately 7 million letters in my large 'dictionary' (compared with, for example, the 10 thousand letters of running text, not dictionary words, of Helen Gaines in *Cryptanalysis*). Rather than list numbers, and to avoid complex graphs, I chose the method shown, which I trust shows rank clearly, but with a strong indication also of numerical value. To take an example, in words of length four, the letters B,C,G,K and W each occur between 2.51% and 3% of the time.

Unfortunately the table has been split in order to fit the page; even so, the steadily increasing importance of O is very clear, to the point where it dominates in longer words. Indeed, the variation with word length is evident. Despite this, the frequency of occurrence of the five vowels as a whole (rightmost column) is rather constant at around 37% to 40% (except for the always peculiar two-letter "words"). Differences in frequency have been noted in Card and Eckler, "A Survey of Letter-Frequencies" in the May 1975 Word Ways, but this presentation shows the consistent change with word length. Some letters, like M, have a remarkably constant frequency, in this case just below the value of 3.8% to be expected if every letter were equally frequent.

The last few rows give the overall figures ('Dict'), the figures for the number of letter tiles of each letter as a percentage of the 98 letter tiles in Scrabble ('Scrabble'), the figures for initial letters ('Init') in the dictionaries, and the figures for terminal letters ('Ter') in the dictionaries. "Ter" includes plurals. The Scrabble figures follow common usage reasonably, though the maximum of 12 of any letter (just E), and the fact that no letter has 7 or 11 tiles, means that infrequent letters are overrepresented, and gradations in frequency are coarser than necessary. Scrabble letters also have a points value, ranging from 1 for AEILNORSTU to 8 for JX and 10 for QZ (with values 6,7 and 9 missing). These values are also related to scarcity of occurrence in words.

Length / %	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5
2			Z		J	KQX	GW	BFHRY	CLMNPTUV	ADEIO	S						
3	Q	J	XZ		KV	FW	BGH	LMN	CDPRY	TU		S	I	O		E	
4	Q	JXZ	V		F	BCGKW	HMP	D	Y	LNU	T	R	IS		O		
5	JQX	Z	V	F	BGKW	MP	CDH	Y	U		NT	L	IR	O	S		
6	JQXZ	V	F	KW	B	GHPY	CM	D	U			LNT	O	IS	R		
7	JQXZ	V	FKW		BY	GHP	M	CDU				LT	NO		IRS		
8	JQXZ	V	FKW	Y	B	GHP	M	CDU				L	NT	O	RS	I	
9	JQXZ	V	FKW	BY	G	H	MPU	D	C			L	T	NO	RS		I
10	JQXZ	KVW	F	BY	G	H	MPU	D	C			L		NT	ORS		A
11	JQXZ	KVW	F	BY	G	H	DMPU			C		L		NST	OR		A
12	JQXZFKVW			B	GY		DHMU	P			C	L		NT	RS	O	A
13	JKQWXZ	FV	B		GY	D	HMU	P			C	L		N	RST		AO
14	JKQWXZ	FV	B		GY	DU	HM	P			C		L	NR	ST		AO
15	JKQWXZ	FV	B		DG	UY	HM	P				C	LN	R	ST		A
16	JKQWXZ	FV	B	G	D	U	HMY		P			C	N	LR	ST		A
17	JKQWXZ	FV	B		DGU		MY	H	P			N	C	LRS	T		A
18	JKQWXZ	FV	B		DGU		MY		HP		N		CS	LR	T		A
19	FJKQVWX	BZ		U	DG			MY		HNP		S	CR	L		T	A
20	FJKQVWX	BZ		U	DG			MY		HNP		S	C	LR	T		AI
21	FJKQVWX	BZ		U	DG			M	Y	HNP	S		C	LR	T	A	I
22	FJKQVWX	BZ		U	DG		M		Y	HNP	S			CLR	T	A	I
23	FJKQVWX	BZ		U	D	G		M	Y	NP	HS		C	LR	T	AI	
24	FJKQVWX	BZ		U	D	G	M	Y		NP	HS		L	CRT		AI	
25	FJKQVWX	BZ			U	DG	M		Y	NP	H	S	L	CRT	I	A	E
26	FJKQVWX	BZ			GU	D	M		Y	HPS		N	C	ILRT	A		
27	FJKQVWX	BZ			DGU			MY	HNS	P		CL	I		AR		
28	FKQVWX	BJZ		DU	G	Y		M	P	S	HN	CL		A	IR	T	
29	BFKQVWX	J	Z		DGU		MY	S			P	HR	CNT	L	AI		
Dict	JQXZ	KVW	F	B	GY	H	DMPU			C		L	N	T	RS	O	A
Corpus	JQXZ	K	V	BPY	FGMW	CU			DL			HR	S	N	I	O	A
Common									H		L	IS	R	N	A	T	O
Scrabble			JKQXZ		(rem)		G		DLSU				NRT				O
Init	XZ	JQY	K	V	NW	O	FGILU	E	HR	D	B	MT			A		
Ter	BFJQUVUXZ	IOP	K	H	M		C		AG	L		RT		N			D

I wanted to compare my results with figures for running text. I am grateful to Ramesh Krishnamurthy of COBUILD, University of Birmingham, England, for permission to use his unpublished research on the COBUILD corpus. At the time of the study, the corpus consisted of almost 17 million words from newspapers, magazines and books, with 75% from the UK, 20% from the US and 5% from South Africa, Australia, etc. The texts are all modern (within the last three decades or so). The corpus now has 300 million words. The results are on the line starting 'Corpus'. They agree excellently with the traditional order found in typesetting shops (ETAOIN SHRDLU). This may show that typesetters know better than the authors of some published frequencies based on small samples; such studies give similar but less accurate results. Of course, one might expect some variation with type of text.

Card and Eckler also looked at the letters in the 500 commonest words (according to Kucera and Francis, Computational Analysis of Present-Day American English, Brown University Press, a corpus of about one million words, and as such I believe containing fewer letters than the dictionary I used!). Their results for letters in common words are in fair agreement with the COBUILD corpus; the top few letters are shown on

Length / %	9	9.5	10	10.5	11	11.5	12	12.5	13	13.5	14	14.5	15	15.5	16	Vowels
2																23.13
3	A															34.36
4			A	E												37.76
5		A					E									38.03
6	A								E							38.33
7	A						E									38.39
8	A						E									38.66
9	A						E									38.81
10	I					E										38.88
11		I			E											38.88
12		I		E												38.87
13		I		E												38.92
14		I	E													38.79
15	O	I	E													38.81
16		EIO														38.76
17		EI	O													38.63
18	E	I		O												38.51
19	EI				O											38.37
20	E					O										38.38
21	E						O									38.36
22	E							O								38.36
23		E						O								38.23
24		E						O								38.54
25									O							37.86
26			E							O						38.84
27	T			E						O						39.87
28						E					O					40.13
29			E									O				40.47
Dict	I				E											38.71
Corpus		T						E								38.08
Common										E						38.00
Scrabble		AI		E												39.84
Init		C	P		S											20.66
Ter		Y													E = 16.46	S = 21.43 22.69
																22.69

the line starting 'Common'.

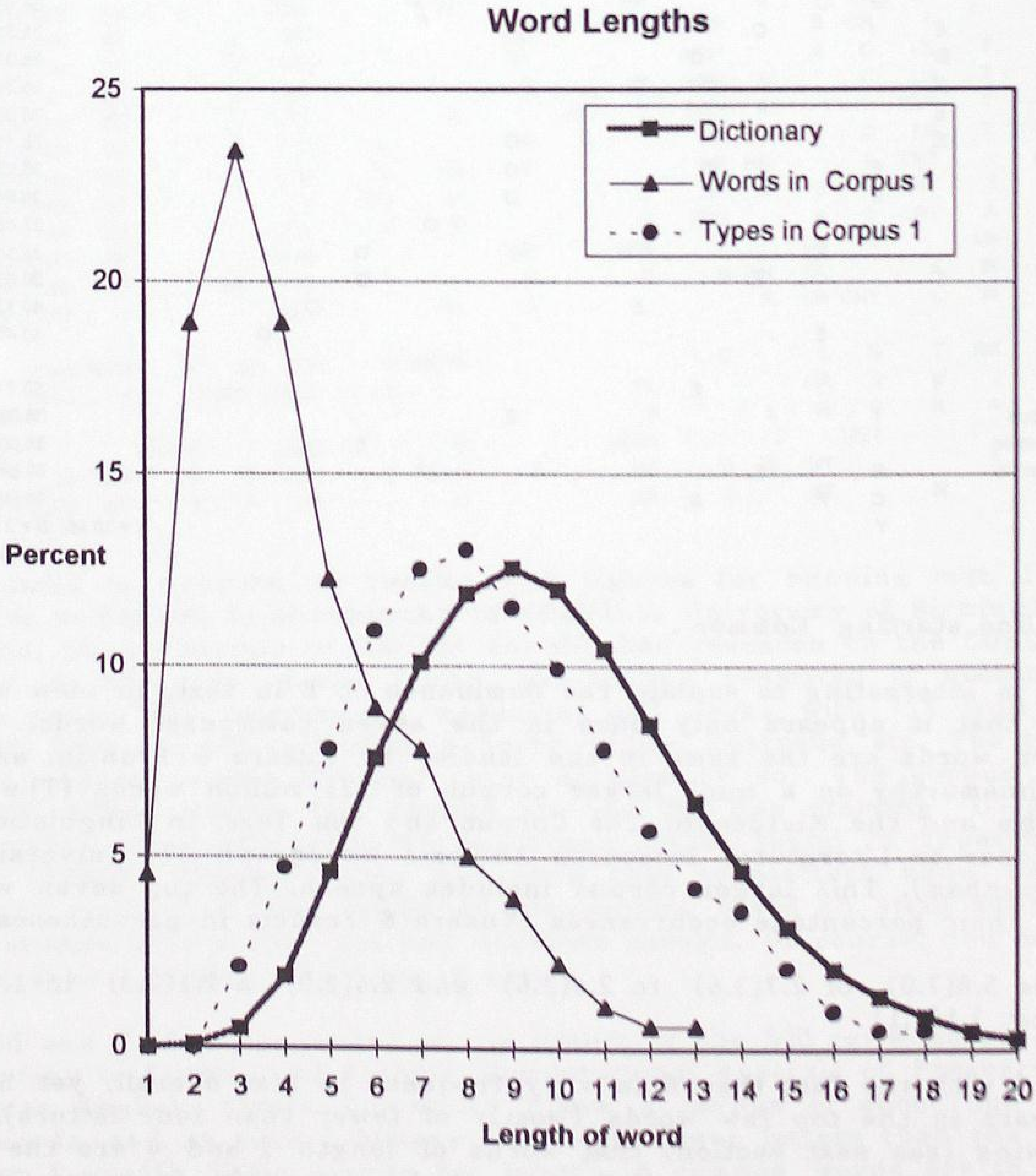
It is interesting to explain the dominance of E in text, in view of the fact that it appears only once in the seven commonest words. These seven words are the same in the studies by Kucera & Francis, and by Krishnamurthy on a much larger corpus of 121 million words (The Macrocosm and the Microcosm: The Corpus and the Text, in Linguistic Approaches to Literature, Discourse Analysis Monograph 17, University of Birmingham). This larger corpus includes speech. The top seven words, with their percentage occurrences (Kucera & Francis in parentheses) are

> the 5.8(7.0) of 2.7(3.6) to 2.6(2.6) and 2.4(2.9) a 2.1(2.3) in 1.9(2.1)
> that 1.1(1.1)

Faced with the fact that E is very frequent in text overall, yet hardly appears in the top few words (mostly of fewer than four letters), and knowing (see next section) that words of length 3 and 4 are the most common, we might guess that E must be extremely frequent in the common words of four letters or more. Indeed, the table confirms this: E becomes the most common letter in four-letter words, then becomes increasingly dominant as the word length increases, reaching a peak at six-letter words, but remaining dominant until we reach the rare length of 17 letters. The results of Card and Eckler for the letters in the 500 commonest words are compatible with this explanation.

WORD LENGTHS

In this section, much of the basic information is taken from the aforementioned monograph of Krishnamurthy, and may be slightly in error because it has been read off rather small graphs. The source of the data for two of the graphs is the 17 million word corpus mentioned before (Corpus 1).



There are 220,000 different words among the 17 million (in the jargon, 220,000 types). The frequency of types corresponds very well to my much larger list of words, as the graph shows, except that my graph is shifted to the right. The reason for this is that the corpus has a smaller percentage of the long words that exist than of the short words, and this is at least partly due to the fact that it contains little of a technical nature (few chemicals or -ostomies!). The fact that the dictionary contains some much longer words is not numerically significant.

The difference between these two graphs and the other is striking, and is due to the many repetitions of the short popular words THE, OF, TO, AND, A, etc. The long right-hand tail of this graph is the reason why the arithmetic average of 4.9 letters per word is a misleading measure of centrality to use: clearly 3 is the commonest word length (the mode). It is clear that the average length of the types, and of the dictionary words, exceeds 8, though this merely serves to emphasize that long words are used little in relation to their number.

Krishnamurthy makes the interesting point that, both in Corpus 1 and in the 121-million word Corpus 2, half of all the types occur only once. He also observes that speech (included in Corpus 2) has the shortest average word length, followed by newspapers, then books. Some would say that more formal documents would tend to use longer Latin types in preference to Anglo-Saxon. In addition, the shortest words have an affinity for each other, the commonest pairs of words being OF THE, IN THE, TO THE, FOR THE, ON THE, TO BE, and AT THE, so the short words reinforce their dominance in text.