Behavior Research Methods & Instrumentation
1976, Vol. 8 (3), 283-286

Frequency and versatility of letters in the English language

ROBERT L. SOLSO
University of Idaho, Moscow, Idaho 83843

and

JOSEPH F. KING
Loyola University of Chicago, Chicago, Illinois 60626

Tabulations of letter and letter-combination versatility and frequency were made based on the Kucera and Francis (1967) word frequency count. Letter versatility, a new descriptive statistic, was defined as the number of different words in which a letter appears. These tabulations may be useful in the investigation of visual information processing, reading skills, and human memory.

A powerful instrument in the understanding of verbal processing has been the statistical analysis of the language. These descriptive statistics can be traced to the 14th century, when Simonetta (as reported by Cherry, 1966) published letter frequencies that were used in cryptoanalysis, Since then, language has been statistically analyzed in terms of bigram and trigram frequency, vowel and dipthong frequency, syllabic frequency, and other variables (see Cherry, 1966; Kahn, 1967; Miller, 1951; Underwood & Schulz, 1960, for more complete historical reviews).

A significant methodological change from counting raw frequencies of letters was introduced by Mayzner and Tresselt (1965); single-letter frequencies were measured by the position each letter occupied in words of varying length. Positional frequency has been hypothesized by Mason (1975) to augment visual feature information in the identification of letters. Spatial redundancy in her study was defined as "the degree of correlation present in printed English between visual features (fixed levels on physical dimension) and their spatial location within multiletter configuration." By having sixth grade readers search through single six-letter strings for either the presence or absence of a specific letter target, she demonstrated that *spatial* redundancy rather than familiarity is the relevant redundancy variable. However, the concept of versatility was not introduced.

In a recent publication (Topper, Macey, & Solso, 1973), bigram frequency and a new descriptive statistic, bigram versatility, were reported. Bigram versatility is defined as the number of different words in which a given bigram appears. Bigram frequency, on the other hand, is the total number of occurrences of a bigram in the language sample. The potency of bigram versatility as it relates to memory search time has been demonstrated by Solso, Topper, and Macey (1974).

In the present paper, a definitive count of single-letter frequency and versatility is reported. This information may be of primary interest to those researchers investigating visual information processing, reading skills, memory search, verbal learning, and related areas.

A portion of this paper was presented at the annual meeting of the Psychonomic Society, Denver, November 1975. The authors thank Lin Morris for her suggestions on the manuscript and composition of the figure. Reprint requests should be sent to Robert L. Solso, Department of Psychology, University of Idaho, Moscow, Idaho 83843.

METHOD

All frequency and versatility counts are based on the Kucĕra and Francis (1967) frequency count of about one-million words in the English language. All hyphenated words, words containing apostrophies, and numbers were excluded from the analysis. Approximately 40,000 different words with an approximate total frequency of 1,000,000 occurrences were used.

The procedure for counting total frequency was as follows: The word NOTE has a frequency of 127 per million. The frequency totals for the letters "N" "O" "T" "E" are incremented by 127. The word ATTACK has a frequency of 105 per million: "A" is incremented by 210 as is "T," while "C" and "K" are increased by 105. Since versatility represents the number of different words in which a particular letter combination occurs, the versatility of "A," "T," "C," and "K" increase by one.

The normative data in this report are confined to total frequency and versatility for all words in the Kucera and Francis count and to letter and frequency counts by position for four- and five-letter words. In Table 1, the total single-letter frequencies and versatilities are represented, with percentage of usage based on total usage in parentheses. The ratio between frequency and versatility is also shown in Table 1 (FV ratio), as are the FV ratios based on percentage usage in parentheses.

Tables 2 and 3 contain frequency counts by position for four-

Table 1
Single-Letter Frequency and Versatility (and Percentage) Occurrence in the English Language

Letter	Frequency	Versatility	FV Ratio
A	348411 (7.61)	20372 (7.80)	17.10 (.97)
B	70714 (1.54)	5478 (2.09)	12.91 (.74)
C	142336 (3.11)	11279 (4.32)	12.62 (.72)
D	181054 (3.95)	11140 (4.26)	16.32 (.93)
E	577583 (12.62)	25794 (9.88)	22.39 (1.28)
F	107219 (2.34)	3830 (1.46)	27.99 (1.60)
G	89499 (1.95)	8043 (3.08)	11.13 (.63)
H	252191 (5.51)	7017 (2.68)	35.94 (2.05)
I	336166 (7.34)	20587 (7.88)	16.33 (.93)
J	7296 (.15)	649 (.24)	11.24 (.63)
K	29838 (.65)	2942 (1.12)	10.14 (.58)
L	188261 (4.11)	14259 (5.46)	13.20 (.75)
M	116574 (2.54)	7975 (3.05)	14.62 (.83)
N	325652 (7.11)	18768 (7.19)	17.35 (.99)
O	350121 (7.65)	16050 (6.14)	20.29 (1.25)
P	93040 (2.03)	7756 (2.97)	12.00 (.68)
Q	4936 (.10)	545 (.20)	9.06 (.50)
R	281881 (6.15)	19243 (7.37)	14.65 (.83)
S	297531 (6.50)	18977 (7.27)	15.68 (.89)
T	427179 (9.33)	17257 (6.61)	24.76 (1.41)
U	124736 (2.72)	9665 (3.70)	12.91 (.73)
V	45707 (.99)	3423 (1.31)	13.35 (.75)
W	86563 (1.89)	2769 (1.06)	31.26 (1.78)
X	9032 (.19)	847 (.32)	10.66 (.59)
Y	78749 (1.72)	5344 (2.04)	14.74 (.84)
Z	4316 (.09)	973 (.37)	4.44 (.24)

Table 2
Letter Frequency and Versatility by Position for Four-Letter Words

	Position									
	1		2		3		4		Totals	
Letter	F	V	F	V	F	V	F	V	F	V
A	3210	108	27208	468	20437	205	1141	98	51996	879
B	7428	175	255	11	319	42	343	35	8345	263
C	4726	146	100	12	6748	87	102	17	11676	262
D	4720	132	408	19	2567	78	10212	142	17907	371
E	3774	65	21088	307	22051	204	33108	347	80021	923
F	10660	119	4	3	1671	29	563	27	12898	178
G	3243	101	136	10	1236	59	1497	57	6112	227
H	9913	128	30082	88	420	19	12353	66	52768	301
I	2218	36	22376	302	9800	162	94	45	34488	545
J	2153	56	2	1	12	4	1	1	2168	62
K	2327	60	93	11	3096	37	5658	126	11174	234
L	7988	139	2973	111	11378	191	7456	124	29795	565
M	10932	130	72	19	6969	81	8437	72	26410	302
N	2185	60	5785	45	8459	231	14394	156	30823	492
O	4313	63	24409	420	10904	146	3294	74	42920	703
P	3625	142	981	16	1086	55	1779	82	7471	295
Q	20	5	1	1	0	0	4	2	25	8
R	3192	115	6703	125	13867	209	5415	83	29177	532
S	10963	203	757	15	10248	146	11847	346	33815	710
T	30738	141	578	24	12350	122	25084	218	68750	505
U	1349	15	7307	234	2840	78	75	26	11571	353
V	1357	43	2848	12	5754	40	11	5	9970	100
W	22441	95	575	24	1567	40	2173	31	26756	190
X	1	1	70	7	479	10	100	17	650	35
Y	2028	27	766	29	1010	25	10295	104	14099	185
Z	77	12	4	3	313	17	145	16	539	48
	155581	2317	155581	2317	155581	2317	155581	2317	622324	

Table 3
Letter Frequency and Versatility by Position For Five-Letter Words

	Position											
	1		2		3		4		5		Totals	
Letter	F	V	F	V	F	V	F	V	F	V	F	V
A	7698	239	11407	725	12769	451	5830	267	904	178	38608	1860
B	5586	325	2166	16	947	96	524	61	49	13	9272	511
C	6733	327	582	45	1052	116	8548	186	624	40	17539	714
D	2604	210	676	44	2800	114	3147	165	12022	253	21249	786
E	2290	104	10330	501	15037	285	19236	764	21890	603	68783	2257
F	5856	222	1531	11	368	41	497	45	527	32	8779	351
G	3824	175	714	14	3953	119	1578	115	3391	48	13460	471
H	3592	175	18960	189	1783	33	2541	48	7236	160	34112	605
I	979	57	10353	424	13954	380	7200	271	172	55	32658	1187
J	626	66	2	2	299	8	6	3	0	0	933	79
K	761	82	106	16	1220	46	1397	135	2283	108	5767	387
L	4471	207	5021	274	3175	297	12004	282	5122	168	29793	1228
M	4513	225	1410	38	2303	140	1818	121	488	55	10532	579
N	2415	88	3093	109	4764	308	8837	295	5687	269	24796	1069
O	3171	75	19107	625	10986	329	2952	212	661	115	36877	1356
P	4687	242	1329	63	1143	77	989	105	792	52	8940	539
Q	559	20	118	6	4	3	2	1	0	0	643	30
R	3219	176	8173	368	8999	374	9474	231	12029	244	41894	1393
S	11986	507	995	40	2290	176	8270	199	15947	936	39488	1858
T	14220	245	5851	101	4968	192	6444	305	11311	267	42794	1110
U	2103	42	3415	319	9489	211	3283	113	18	9	18308	694
V	1033	75	697	17	3087	96	1305	53	10	2	6132	243
W	12761	141	716	49	805	56	1144	42	408	17	15834	305
X	11	4	163	13	406	32	10	3	176	16	766	68
Y	1587	25	368	52	551	56	128	19	5523	410	8157	562
Z	17	12	19	5	150	30	138	25	32	16	356	88
	107302	4066	107302	4066	107303	4066	107302	4066	107302	4066	536510	

and five-letter words. In Table 2, 2,317 four-letter words were found which have a total frequency of 155,581. F and V positional percentages are obtained by dividing the specific letter/position by the sum of the column F or V. Table 3 is based on 4,066 words with a total frequency of 107,302.

Complete normative data on words of different length and larger letter combinations may be found in Solso (Note 1). Included are bigram frequencies and versatility, trigram frequencies and versatility, and positional norms.

The relationship between letter frequency and versatility is represented in Figure 1. In Figure 1, frequency by rank order is along the abscissa, while percentage occurrence of letters by versatility and frequency is represented as the ordinate. The general covariation of these two measures is apparent, as are anomalies (e.g., H, which enjoys high frequency due to "the," "there," "this," etc., but relatively low versatility).

The frequency of encountering letters in written material is an overall measure of familiarity, but the versatility values reported here represent the frequency of familiarity of the context within which a letter appears in a word.

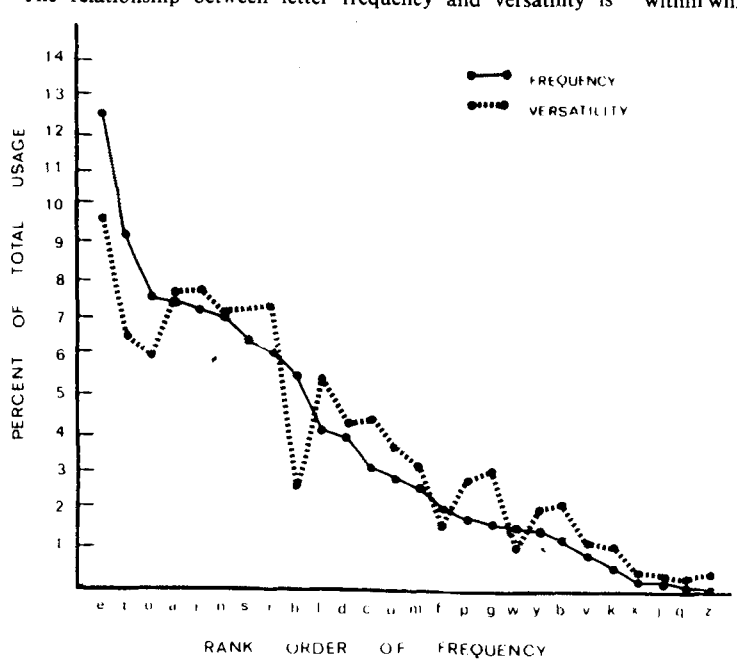

Figure 1. Total usage of English letters by rank order frequency and versatility.

REFERENCE NOTE

1. Solso, R. L. *Statistical analysis of letters and letter combinations in the English language.* Unpublished manuscript.

REFERENCES

CHERRY, C. *On Human Communication.* Cambridge: M.I.T. Press, 1970.

KUCERA, H., & FRANCIS, W. N. *Computational analysis of present-day American English.* Providence: Brown University Press, 1967.

MASON, M. Reading ability and letter search time: Effects of orthographic structure defined by single-letter positional frequency. *Journal of Experimental Psychology: General,* 1975, **1**, 146-166.

MAYZNER, M. S., & TRESSELT, M. E. Table of single-letter and bigram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements,* 1965, **1**(Whole No. 2).

MILLER, G. A. *Language and communication.* New York: McGraw-Hill, 1951.

SOLSO, R. L., TOPPER, G. E., & MACEY, W. H. Anagram solution as a function of bigram versatility. *Journal of Experimental Psychology,* 1973, **100**, 259-262.

TOPPER, G. E., MACEY, W. H., & SOLSO, R. L. Bigram versatility and bigram frequency. *Behavior Research Methods & Instrumentation,* 1973, **5**, 51-53.

UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning.* Chicago: Lippencott, 1960.

(Received November 29, 1975;
revision accepted February 10, 1976.)