

METHODS & DESIGNS

Bigram and trigram frequencies and versatilities in the English language

ROBERT L. SOLSO and PAUL F. BARBUTO, JR.
University of Idaho, Moscow, Idaho 83843

and

CONNIE L. JUEL
University of Texas, Austin, Texas 78712

A comprehensive count of bigram and trigram frequencies and versatilities was tabulated for words recorded by Kučera and Francis. Totals of 577 different bigrams and 6,140 different trigrams were found. Their frequencies of occurrence and the number of different words in which they appeared are reported in this article.

In recent years the orthographic components involved in letter and word identification have received considerable attention. An important component in some studies has been the positional frequency of letters in the English language (e.g., Mason, 1975); other studies have found that versatility (the number of different words in which letters or letter combinations occur) has affected word processing (e.g., Solso, Topper, & Macey, 1973).

The current research in letter/word processing has required more comprehensive and contemporary measures of the statistical properties of the English language. In the past, several researchers have counted single-letter frequencies (Lysing, 1936; Underwood & Schulz, 1960), bigrams (Underwood & Schulz, 1960; Mayzner & Tresselt, 1965; Topper, Macey, & Solso, 1973), and trigrams (Underwood & Schulz, 1960). These counts, however, have been based on a relatively small sample size and done by hand. Recently, Solso and King (1976) tabulated single-letter frequencies and versatilities based on the Kučera and Francis (1967) norms, which contain about one million words in the English language. The positional frequencies and versatilities have also been recently tabulated for four- and five-letter words (Solso & King, 1976) and for six-, seven-, and eight-letter words (Solso, 1979) based on the Kučera and Francis norms.

In the present paper a comprehensive count of bigram and trigram frequencies and versatilities is reported. It is anticipated that these data will be of interest to those investigators interested in visual information processing, reading skills, memory search, verbal learning, cognitive models, and related areas.

METHOD

All bigram and trigram frequency and versatility counts were based on the Kučera and Francis (1967) tabulation of about one million words in the English language. The Kučera and Francis words are stored on magnetic tape, and the bigram and trigram counts reported in this paper were processed by means of an IBM 370/145 computer. Approximately 40,000 different words with about a total frequency of one million occurrences were used. In the present analysis, all hyphenated words, words containing apostrophes, and numbers were excluded.

The procedure for counting total bigrams and trigrams was as follows: The word "root" has a frequency of 30/million. The frequency of the bigrams RO, OO, and OT were all increased by 30. In counting, trigrams ROO and OOT were also increased by 30. For the versatility count, the above bigrams and trigrams were increased by one. The word "cocoon" has a frequency of 3/million. In this case the frequency of bigram CO was increased by six, as it appears twice in the word, but the versatility count was increased by only one. The other bigram and trigram frequencies were incremented by three and versatilities by one. The word "ringing" has a frequency of 10/million. In this case the frequency of the bigrams IN and NG and the trigram ING were incremented by 20, and the versatilities incremented by 1. The other bigram and trigram frequencies were incremented by 10 and versatilities by 1. A total of 577 bigrams (out of a possible 676) were found in the sample. The total bigram frequency as determined by summing all bigram frequencies was 3,616,085. The total bigram versatility as determined by summing all bigram versatility was 270,337. However, because these data represent repeated bigram counts in words, a more practical use of the data may be to use the frequency and versatility counts in relation to the total words and total different words in the Kučera and Francis (1967) count. Thus the bigram EX appeared 6,847 times in about one million words and appeared in 556 different words out of about 40,000 words.

A total of 6,140 trigrams (out of a possible 17,574) were found in the sample. The total trigram frequency as determined by summing all trigram frequencies was 2,662,964. The total trigram versatilities was 233,420. As with bigram interpretation, it may be more useful to use the base of a total of one million words and 40,000 different words. Thus, the trigram HRO appeared 1,490 times in about one million words and appeared in 91 different words out of 40,000 words.

The results of total frequency and versatility counts for bigrams and trigrams are shown in Tables 1 and 2.

Paul F. Barbuto, Jr., is currently at Texas Instruments, Dallas, Texas 75260.

Table 2 (Continued)

Table with multiple columns and rows containing alphanumeric codes (e.g., PKE, PLA, PLE, PLO, PLU, PLY, PMA, PME, PMD, PNE, PNI, PNO, POA, POB, POC, PDD, PDE, PDF, POG, POH, POI, POK, POL, POM, PON, PPO, PPR, PPS, PPU, PRA, PRI, PRO, PRU, PSA, PSC, PSE, PSG, PSH, PSI, PSK, PSM, PSN, PSO, PST, PSU, PSW, PTA, PTC, PTE) and numerical values.

Table 2 (Continued)

Table with 28 columns (A-Z) and 28 rows (A-Z). Each cell contains a count for a specific bigram or trigram combination. The table is a grid of 28x28 cells, with the first row starting with 'SKS' and the first column starting with 'F'. The counts range from 1 to 418.

Table 2 (Continued)

	F	V		F	V		F	V		F	V
XTH	293	40	YGA	1	1	YPL	1	1	ZEM	14	4
XTS	8	3	YGE	46	2	YFN	4	4	ZEN	260	21
XTU	66	9	YGI	4	2	YFO	109	26	ZER	108	35
XTY	21	1	YGM	3	1	YFR	16	5	ZES	99	4
XUA	75	1	YGN	1	1	YFS	10	5	ZET	11	1
XUB	11	6	YGO	12	1	YFT	33	18	ZFY	11	1
XUD	4	3	YGR	4	1	YFA	65	18	ZFE	6	1
XUL	4	2	YGU	1	1	YRD	10	2	ZGE	1	1
XUR	36	5	YGY	1	1	YRE	20	7	ZHA	1	1
XVI	1	1	YHE	1	1	YRI	65	22	ZHE	1	1
XWE	11	1	YHO	47	13	YRN	2	1	ZHI	2	2
XWO	2	2	YHR	1	1	YRC	179	29	ZHO	1	1
XXX	2	1	YHY	1	1	YRS	2	1	ZHU	1	1
XYO	2	1	YKI	4	1	YRT	1	1	ZIA	4	2
XYG	46	2	YKO	64	5	YRV	1	1	ZIB	5	2
XYH	1	1	YIN	1182	136	YRU	15	4	ZIC	3	2
XYL	6	3	YIP	1	1	YRZ	1	1	ZID	2	1
XYM	1	1	YIS	10	7	YSA	1	1	ZIE	42	13
XYT	3	2	YJA	2	2	YSD	2	2	ZIF	2	1
YAB	22	7	YKA	17	2	YSE	2	2	ZIG	6	6
YAC	16	9	YKE	10	4	YSF	16	4	ZIH	19	4
YAD	25	5	YKI	5	1	YSG	44	13	ZIM	4	2
YAG	25	5	YKN	2	1	YSL	3	3	ZIN	33	9
YAH	1	1	YLA	53	18	YSM	3	2	ZIO	12	3
YAJ	1	1	YLB	1	1	YSO	8	6	ZIP	3	3
YAK	4	4	YLE	155	20	YSP	2	2	ZIR	1	1
YAL	153	19	YLG	2	1	YSS	18	5	ZIS	13	2
YAM	5	3	YLI	89	21	YST	33	6	ZIT	3	7
YAN	180	39	YLK	7	1	YSU	33	13	ZKC	7	4
YAP	1	1	YLL	32	12	YTE	3	1	ZKY	4	2
YAU	1	1	YLD	51	15	YTH	629	27	ZLE	77	27
YAR	178	27	YLP	2	2	YTI	63	15	ZLI	26	7
YAS	5	4	YLS	9	5	YTO	37	8	ZLO	1	1
YAT	19	3	YLT	1	1	YTR	1	1	ZLY	2	2
YAU	7	1	YLU	5	2	YTT	7	3	ZMA	2	2
YAW	5	5	YLV	51	5	YUB	3	2	ZME	1	3
YBA	3	3	YLY	15	6	YUC	2	2	ZOA	13	3
YBE	157	10	YMA	39	17	YUG	12	2	ZOD	3	3
YBI	2	1	YMB	165	11	YUH	1	1	ZOE	5	3
YRO	127	10	YME	282	37	YUJ	1	1	ZOG	1	1
YBR	1	4	YMI	7	5	YUK	6	2	ZOI	1	1
YBU	51	7	YMN	22	7	YUN	2	2	ZOL	16	3
YBY	1	1	YMO	36	7	YUR	7	2	ZOM	2	2
YCA	2	2	YMP	75	16	YUS	11	4	ZON	9	18
YCE	46	14	YMQ	209	31	YVA	3	1	ZOO	19	10
YCH	160	34	YMS	5	2	YVB	1	1	ZOP	16	1
YCI	7	3	YMT	2	2	YVI	5	9	ZOR	17	3
YCK	3	3	YAY	2	2	YWA	90	9	ZOS	3	2
YCL	69	21	YNA	68	22	YWE	4	3	ZOT	2	2
YCO	25	17	YNB	9	1	YWH	86	2	ZOU	2	2
YCR	3	1	YNC	20	12	YWI	10	3	ZOV	5	1
YDA	18	3	YND	18	9	YWO	39	4	ZOW	2	1
YDE	10	6	YNE	101	28	YWR	6	2	ZRA	2	1
YDI	8	2	YNG	1	1	YX	1	1	ZRI	1	1
YDN	7	2	YNH	3	1	YZA	8	4	ZRC	1	1
YDO	2	2	YNI	18	5	YZE	40	10	ZSC	2	2
YDR	88	32	YNK	2	2	YZI	8	1	ZUB	2	2
YEA	1675	17	YNN	12	6	ZAA	11	3	ZUE	5	1
YEB	18	5	YND	68	17	ZAB	64	2	ZUK	1	5
YED	513	53	YNE	1	1	ZAC	3	2	ZUR	14	1
YEF	89	2	YNT	48	15	ZAD	3	1	ZUS	1	1
YEG	1	1	YNX	1	1	ZAE	1	1	ZUT	1	1
YEH	3	2	YOC	6	4	ZAG	9	3	ZVE	1	1
YEI	5	2	YOD	4	3	ZAH	1	1	ZVO	2	2
YEK	1	1	YUF	7	5	ZAI	4	1	ZWE	7	1
YEL	121	23	YOG	4	3	ZAK	3	2	ZWK	1	1
YEN	13	1	YOK	21	9	ZAL	20	7	ZWO	2	2
YEP	7	7	YOL	5	2	ZAN	24	14	ZYC	1	1
YER	367	50	YOM	9	1	ZAP	2	2	ZYB	1	7
YES	668	15	YON	454	21	ZAR	67	25	ZYG	1	2
YET	453	11	YOO	4	3	ZAS	4	4	ZYK	1	1
YEV	8	6	YOP	3	3	ZAT	509	84	ZYM	21	3
YEW	2	2	YOR	367	14	ZBE	3	1	ZYN	3	10
YEW	2	2	YOS	6	4	ZBO	4	1	ZZA	38	7
YEZ	1	1	YOT	11	4	ZBZ	20	6	ZZI	10	12
YFA	2	2	YOV	4899	19	ZEB	2	2	ZZL	38	35
YFE	4	1	YGU	13	2	ZEC	6	6	ZZM	1	4
YFF	3	2	YDW	19	2	ZED	748	161	ZZO	15	1
YFI	1	1	YDX	2	1	ZEE	16	2			
YFL	4	3	YPA	4	2	ZEF	1	1			
YFO	8	1	YPE	382	31	ZEI	4	3			
YFR	1	1	YPH	18	9	ZEK	2	2			
YFU	6	4	YPI	98	10	ZEL	12	9			

REFERENCES

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

LYSING, H. *Secret writing, an introduction to cryptograms, ciphers, and codes*. New York: Kemp, 1936.

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).

SOLSO, R. L. Positional frequency and versatility of letters for six-, seven-, and eight-letter English words. *Behavior Research Methods & Instrumentation*, 1979, 11, 355-358.

SOLSO, R. L., & KING, J. Frequency and versatility of letters in the English language. *Behavior Research Methods & Instrumentation*, 1976, 8, 283-286.

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: Lippincott, 1960.

(Received for publication May 10, 1979; revision accepted July 19, 1979.)