HOW MANY WORDS SUPPORT A SQUARE?

A. ROSS ECKLER
Morristown, New Jersey

Select words of n letters at random from a dictionary, one at a time. How many different words must one select before one can form a word square out of n of them? Define the support of a square as the average value of this number, taken over a large number of repeititions of this experiment. The determination of the support is a task well-suited to the computer, which can not only ensure randomness but exhaustively search for possible squares (1 challenge **Word Ways** computer mavens to estimate the support for squares of size 2 through 6).

This article approximately estimates the support for squares of size 2 through 5 by looking at a related problem: what is the commonest word square of size n, where "common" is taken to mean that the rarest word in the square, as measured by Kucera and Francis's *Computational Analysis of Present-Day American English*, has as *high* a number of occurrences as possible? These (minimax) squares are given below, with the number to the right of a word indicating its observed frequency in one million words of American English published in 1962. Can anyone find squares with higher minima?

O	N	6742
N	O	2201

C	A	N	1772
A	R	E	4393
N	E	W	1635

M	O	N	T	H	130
O	P	E	R	A	47
N	E	V	E	R	698
T	R	E	N	D	46
H	A	R	D	Y	42

S	A	M	E	686
A	W	A	Y	456
M	A	D	E	1125
E	Y	E	S	401

The estimated supports for these squares of size 2, 3, 4 and 5 are, respectively, 15, 21, 68, and 455; in other words, NO was the fifteenth most common two-letter in Kucera and Francis (preceded by OF,TO,1N,IS,HE,1T.AS,ON,BE,AT,BY,OR,AN,WE), NEW was the 21st most common three-letter word, EYES the 68th most common four-letter word, and HARDY the 455th most common five-letter word.

Naturally, these are only approximate estimates of the support. An effort was made to refine the estimated support for the three-square by listing the 94 commonest words and forming *all* possible three-squares from them. 1 found 115, but (not being as patient or thorough as a computer) may have overlooked a few. lt is easy to obtain a better estimate of the support by a simple scaling argument:

$$\text{Support} = (\text{Number of words selected})/(\text{Number of squares found})^{1/n}$$

Substituting in the above values, the support for the three-square is

$$94/115^{1/3} = 94/4.86 = 19.33$$

only a little less than the 21 estimated from the commonest square.

The value of knowing the support becomes clear when one considers larger n. If one is to set a computer to the extremely laborious task of examining all possible n-combinations of a set of words to see if an n-square is lurking there, one wishes some degree of assurance that the set is large enough to make the search a success (with high probability). Of course, one cannot know the value of the support in advance of finding the first square, but one can estimate the support by a modest extrapolation of smaller values of n.

Fortunately, past articles from **Word Ways** provide two data points beyond n=5. In the November 1975 **Word Ways**, Doug McIlroy reported an exhaustive search of 9663 seven-letter words and names drawn from Webster's Seventh Collegiate, which resulted in 54 seven-squares (two belatedly noted in the August 1990 Colloquy). Substituting this into the support equation, one derives a value of 5459. In the November 1991 **Word Ways**, Eric Albert chronicled his successful discovery of a single nine-square in Webster's Second Unabridged. According to the Air Force reverse dictionary list based on the same corpus, there are 36419 solid nine-letter entries therein, and this can be taken as an estimate of the support.

Using the supports for three-, five-, seven- and nine-squares, I propose the following tentative support table for all values of n from three through ten:

n	support	log$_e$	d	d^2
3	19.33	2.96	1.68	
4	104	4.64	1.48	-.20
5	455	6.12	1.32	-.16
6	1706	7.44	1.17	-.15
7	5459	8.61	1.02	-.15
8	15285	9.63	0.87	-.15
9	36419	10.50	0.72	-.15
10	74608	11.22		

In his most recent **Word Ways** update (November 1989), Frank Rubin reported that he had placed 94200 ten-letter words or phrases in his database. Applying the support equation, one concludes that a complete examination of this corpus for possible ten-squares ought to uncover several:

$$94200/x^{1/10} = 74608$$
$$x = 10.32$$

As Rubin points out, however, his program is not fast enough to examine all possibilities; he uses heuristics to eliminate unpromising material. For example, each square evaluated uses only words starting with bigrams for which at least 25 such words exist.

This article has focused on the support for single word squares -- those whose vertical words duplicate the horizontal ones. Exactly the same investigations can be made on behalf of double word squares -- those whose vertical words are all different from the horizontal ones. (There are, also, degenerate forms of the double word square which are only required to have at least one vertical word different from any horizontal word, and vice versa.) Less is known because less data have been assembled. In the case of the double three-square, I could discover only five specimens even when using an enlarged stockpile of 132 words, leading to a support of 77.24, approximately four times the support of the single three-square. For the double seven-square, McIlroy was not successful in locating a single one, setting a lower bound of 9663 on the support. However, his computer did find (**Word Ways**, May 1976) 117 double six-squares. None consisted of words solely from the 4060 six-letter words in Webster's Pocket Dictionary, again providing a lower bound to the support. If his stockpile of six-letter words is taken to be 7500 (McIlroy did not give a figure), the estimated support is, in fact, only 3391. This is a little less than twice the corresponding support for the single six-square, suggesting a possible convergence in support between single and double squares as the square size increases. Eric Albert conjectures that for large squares (say, size nine) there actually exist more double squares than single ones -- that is, the support values cross over. The reason that no really large double squares have been discovered, he believes, is simple: double squares are far more thinly scattered in their "space" than corresponding single squares are, making it very much a needle-in-the-haystack proposition. However, with the inexorable march of personal computer power, it should not be too long before his conjecture can be actually put to the test.