ROOM FOR EXPANSION

ENOCH HAGA
Livermore, California

How many words are possible using the 26 letters of the Roman alphabet? I've made some speculative calculations that I believe provide a preliminary answer to this question. Since languages other than English make use of the Roman alphabet, and probably more Romanization will occur in the future, my numerical estimates include their usage. But I will focus on English, and assume that English alone may look to the entire pool of possible words as needed for future expansion.

English is growing at one end with the constant creation of new words, for example *e-mail* (or *email* as I prefer), and dying at the other with the declining use of obsolescent and obsolete language, for example *gay* in the sense of *happy*. Yet even "dead" words must be preserved in dictionary lists, and some may reassume their old meanings. Possibly there can be a word to express each feeling or emotion as well as each discrete thing. How many of these are there? Perhaps the psychologists and biologists can help us out here, for what is the limit of human thoughts and feelings, each of which might need its own word to convey a precise description?

We have in English only 26 letters or symbols, and these are known by linguists to be insufficient to express the range of human sound. Other languages, such as Sanskrit, can make a better fit, with some 40 symbols. Sanskrit, of course, is the root language of English. Chinese, in written form, makes use of characters capable of expressing multiple meanings in a kind of *explosive* rather than *linear* order. The march of subject, verb, and predicate in English sentences shapes our thinking as well as expresses meaning.

As is true of prime numbers, theoretically there is no limit to the potential infinite length of an English word. There are only practical limitations. But will human or extraterrestrial intelligence ever admit common usage of ever-longer words? One measure of intelligence is the length of a string of random numbers that one can accurately repeat back. So as to define the maximum "acceptable" length of new words, this principle might also be applied to strings of random letters.

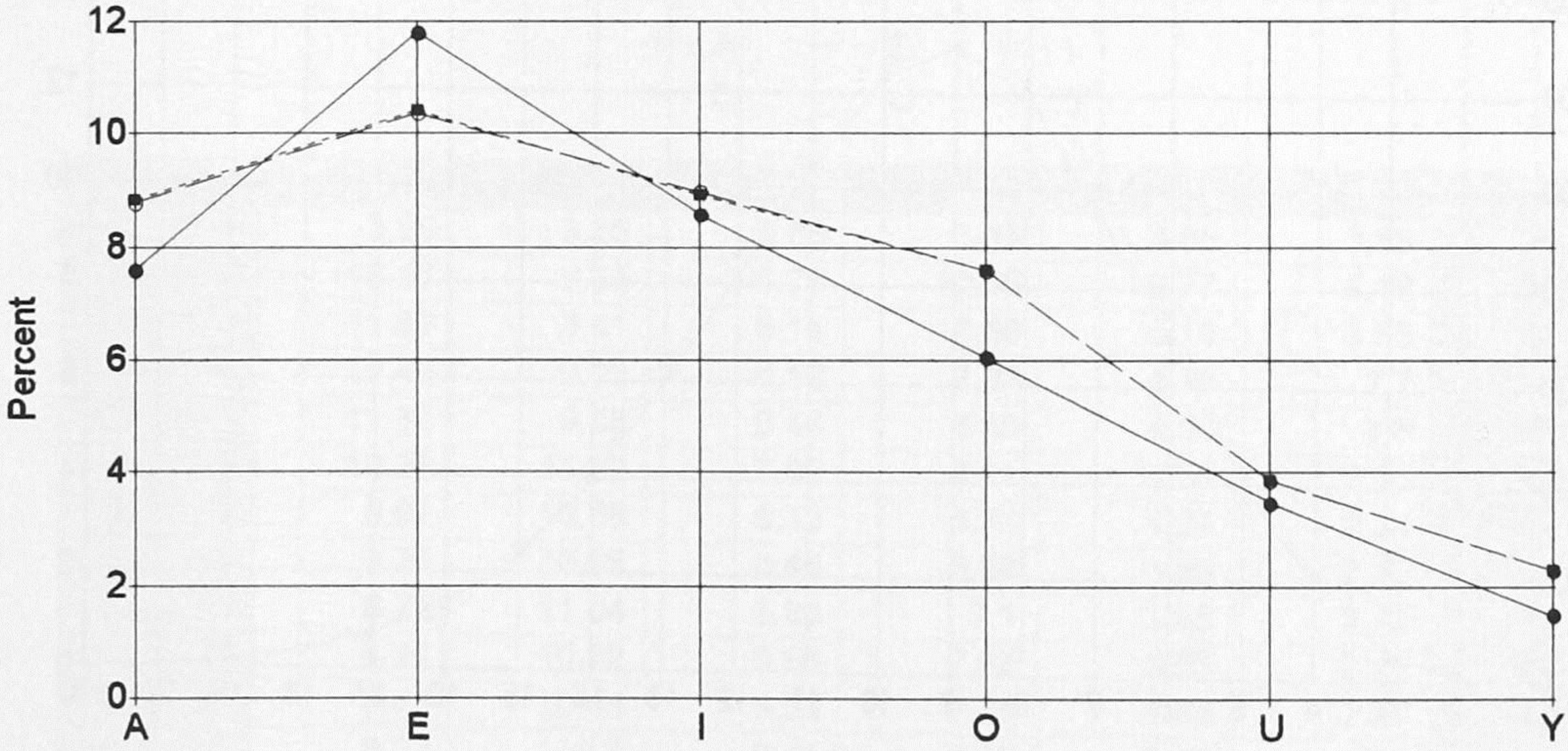
For the present, a length of 20 letters seems to provide a practical working limit. The first step to be taken with reference to English expandability is to locate some large dictionary lists of non-specialized words; these lists may then be analyzed to determine such things as the mean number of vowels in words of various length. Today such lists are readily available for downloading on the Internet (just type "Wordlists" in the blank provided for your search engine). Calculations that once were tedious and time-consuming can now be computed in more detail and in much less time. The work

involved after downloading is unzipping the file, moving the file to a word processor for removal of unneeded codes and punctuation, and saving the file in ASCII format so that the file can be manipulated with one or more computer programs.

English has plenty of room for new words, and vowels must form their backbone. So we'll try to block out the gross word limit in the range of 1 to 20 letters. We'll begin with no prejudice against any specific letter; after all, do we know for certain what specific sound, at some future time, may be assigned to that letter? While new English words must, on average, contain about 58 percent consonants and 42 percent vowels, this is not true of some other languages. Gaines in *Cryptanalysis* (Dover, 1956) notes (p 219) that for French it is 45 per-cent, and for Italian, Spanish, and Portuguese each around 48 percent. The English vowel count of 42 percent is based on actual computer counts in three downloaded dictionary wordlists: Web2, Unabr, and OSPD (*The Official Scrabble Players Dictionary*). As is obvious in the chart below, only Web2 and Unabr are general-purpose dictionaries.

Vowels in 3 Dictionary Wordlists

From Internet Downloads

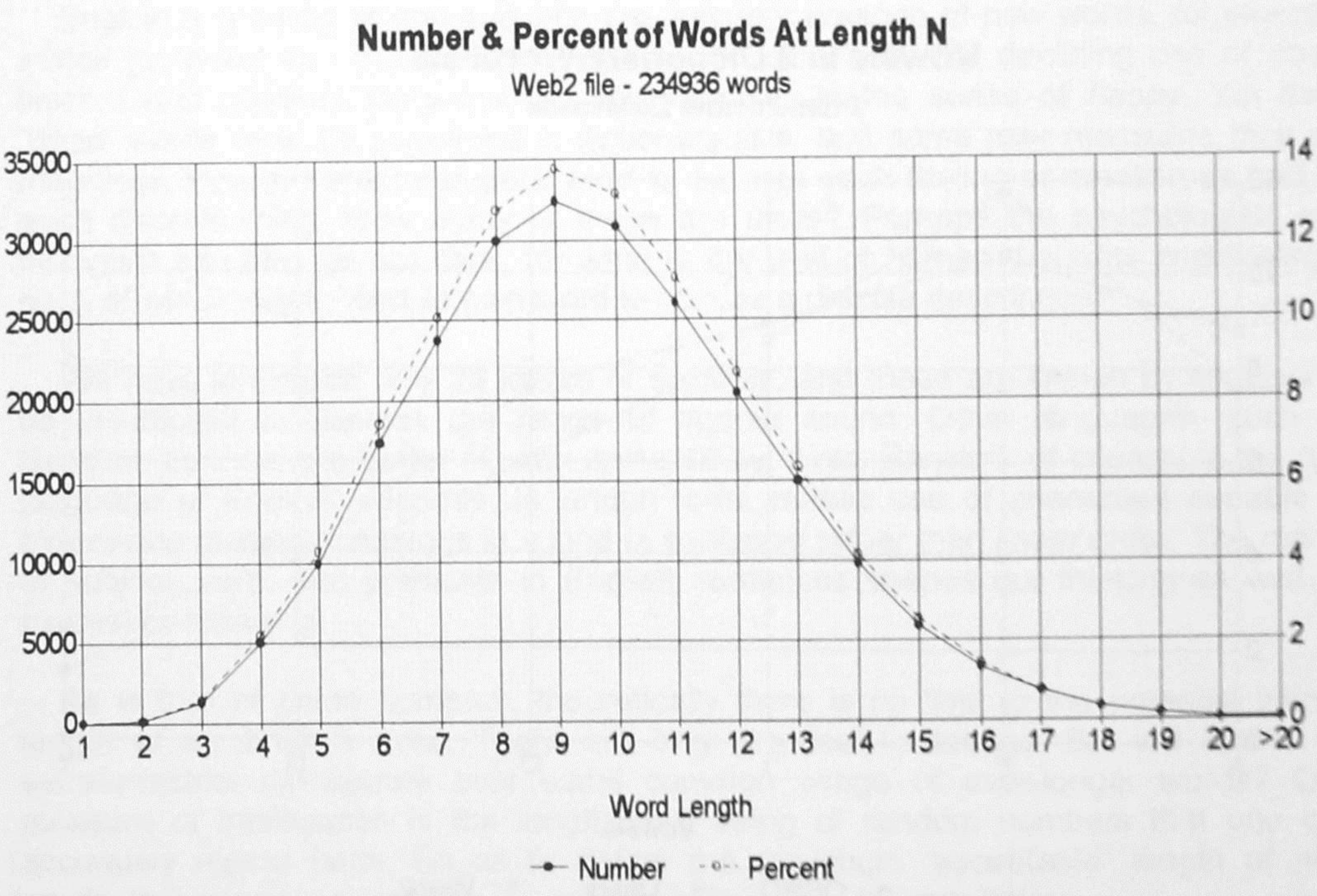


The Unabr and Web2 percents are so close that they cannot be distinguished on the above chart. The OSPD percents reveal the smaller number of words in its list. OSPD has a file of 113801 words and 902847 characters which contain a total 351900 vowels (38.98%). For Unabr there are 213557 words and 2099075 characters containing 880712 vowels (41.96%). For Web2 there are 234936 words and 2251877 characters containing 943346 vowels (41.89%). Based on these calculations, the proportion of

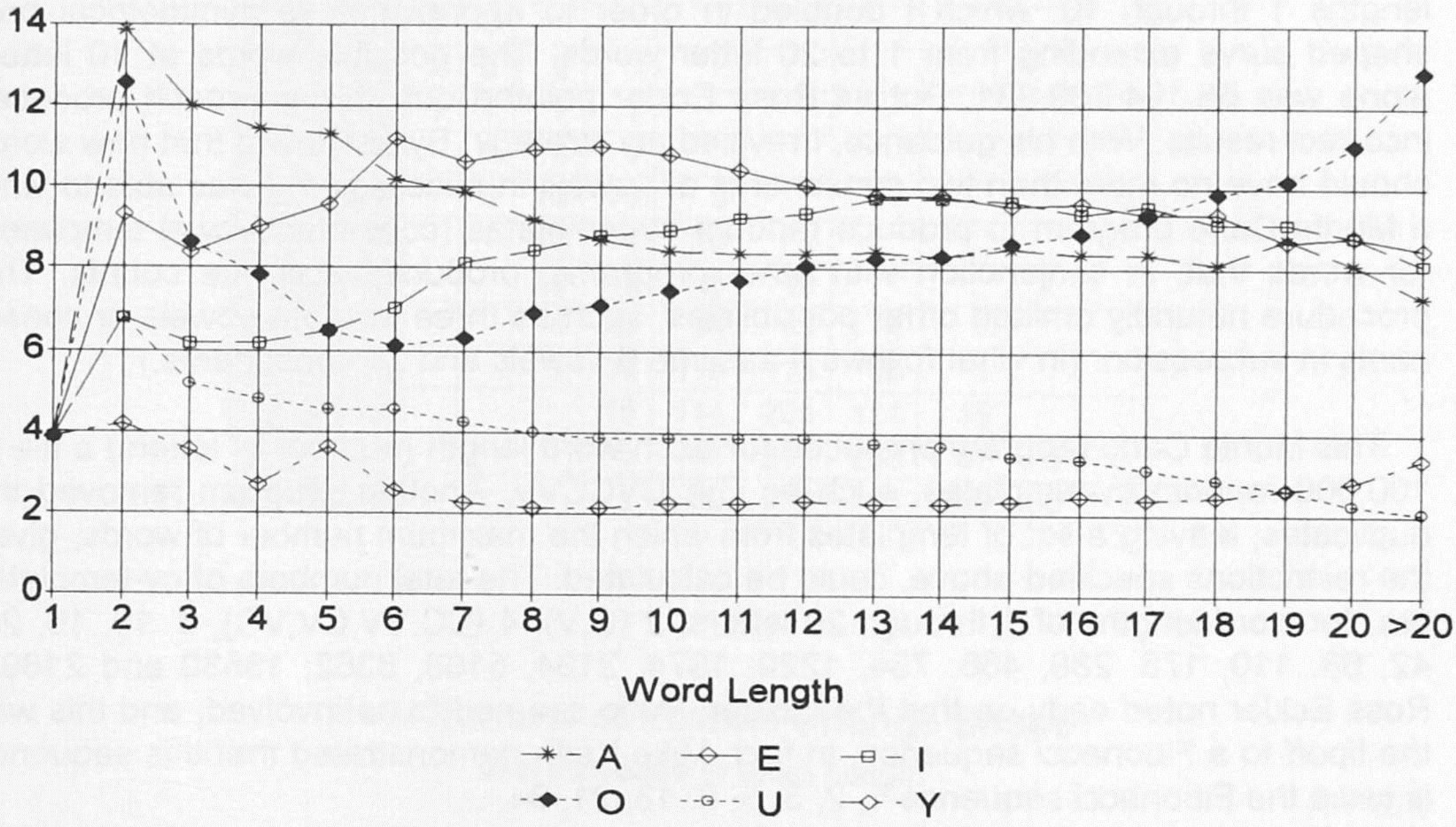
vowels in English dictionary wordlists may be taken as approximately 41.9%, rounded to 42.

The chart and table on the next page, Percent of Vowels, shows for Web2 the percent of vowels found in words of length 1 to 20 letters. Web2 was selected because it has more characters than either OSPD or Unabr. Note the erratic behavior of the vowels at various word lengths. However, as seen in the rightmost column of the table, once past 3 letters, the percent of vowels is remarkably consistent across word length at about 42 percent.

In the chart below for the 234936 words of Web2, we see that the distribution of words of various lengths follows a near-perfect bell-shaped curve (if not truncated at 20 words, the right tail would extend out some distance). The vertical percent scale is to the right, the number scale to the left.



We are now in a position to calculate the theoretical word limit of the English language. We have seen that except for word lengths less than 4, the percent of vowels remains relatively constant at 42 percent. (This is true only of dictionary lists; in actual English writing many words are repeated, effectively lowering the percent to a level closer to 40.) For word length 2 it is 53.5 percent, and for length 3 about 44 percent.



Length	A	E	I	O	U	Y	Total
1	3.85	3.85	3.85	3.85	3.85	3.85	23.08
2	13.87	9.35	6.77	12.58	6.77	4.19	53.54
3	11.99	8.41	6.14	8.66	5.18	3.55	43.94
4	11.43	9.03	6.16	7.83	4.79	2.7	41.94
5	11.35	9.58	6.44	6.49	4.54	3.6	42.01
6	10.23	11.22	7.06	6.11	4.56	2.56	41.74
7	9.87	10.66	8.13	6.31	4.22	2.24	41.42
8	9.21	10.95	8.46	6.93	3.96	2.12	41.64
9	8.72	11.04	8.83	7.1	3.85	2.12	41.65
10	8.46	10.82	8.96	7.48	3.85	2.25	41.82
11	8.43	10.44	9.28	7.71	3.83	2.27	41.96
12	8.37	10.1	9.41	8.06	3.88	2.3	42.12
13	8.45	9.83	9.76	8.29	3.72	2.24	42.28
14	8.36	9.8	9.75	8.36	3.7	2.22	42.2
15	8.5	9.56	9.71	8.65	3.41	2.33	42.18
16	8.41	9.66	9.39	8.88	3.33	2.42	42.1
17	8.38	9.39	9.57	9.31	3.1	2.37	42.08
18	8.16	9.38	9.17	9.91	2.84	2.43	41.9
19	8.8	8.73	9.15	10.21	2.61	2.64	42.14
20	8.13	8.86	8.84	11.06	2.27	2.83	41.99
>20	7.37	8.52	8.17	12.85	2.04	3.38	42.33

Chart and Table for Web2, Percent of Vowels

It is quite easy to produce a computer list of estimated maximum words of any length based on 58 percent consonants and 42 percent vowels (or any other proportion). Using this assumption, I calculated a theoretical total of 71,087,747,273 words of lengths 1 through 10, which I doubled in order to approximate a symmetrical bell-shaped curve extending from 1 to 20 letter words. The possible words at 10 letters alone was 65,194,239,931. But as Ross Eckler pointed out, this approach produced incorrect results. With his guidance, I revised my strategy. By assuming that new words should have no more than two consonants or vowels in succession, I was able to write a Monte Carlo program to produce random cv-templates (consonant/vowel templates) for words that, in conjunction with other programs, produced accurate counts. This procedure naturally omitted other possibilities, such as three or more vowels or consonants in succession. (In what follows, I assume 6 vowels and 20 consonants.)

This Monte Carlo program produced for each word length (number of letters) a file of 100,000 random cv-templates, such as CVCCVCCVV. Another program removed the duplicates, leaving a set of templates from which the maximum number of words, given the restrictions specified above, could be calculated. The total numbers of cv-templates are, for word-lengths of 1 through 20 letters, 2 (C,V), 4 (CC,VV,CV,VC), 6, 10, 16, 26, 42, 68, 110, 178, 288, 466, 754, 1220, 1974, 3194, 5168, 8362, 13530 and 21892. Ross Eckler noted early on that the Golden Ratio seemed to be involved, and this was the tipoff to a Fibonacci sequence. In fact, Mike Keith demonstrated that this sequence is twice the Fibonacci sequence 1, 2, 3, 5, 8, 13, 21, 34, ...

To illustrate the use of the cv-templates for counting vowels, take a word length of 5 as an example. A computer program counts the number of vowels (1 through 4) in the 16 allowable 5-letter cv-templates. Of these 16, one had 1 vowel (CCVCC), seven had 2 vowels, seven had 3 vowels, and one had 4 vowels (VVCVV). (Notice the symmetry of the pattern.) Translating these vowel counts into power functions we obtain:

(20x20x20x20)(6)(1) = 960000
(20x20x20)(6x6)(7) = 2016000
(20x20)(6x6x6)(7) = 604800
(20)(6x6x6x6)(1) = 25920

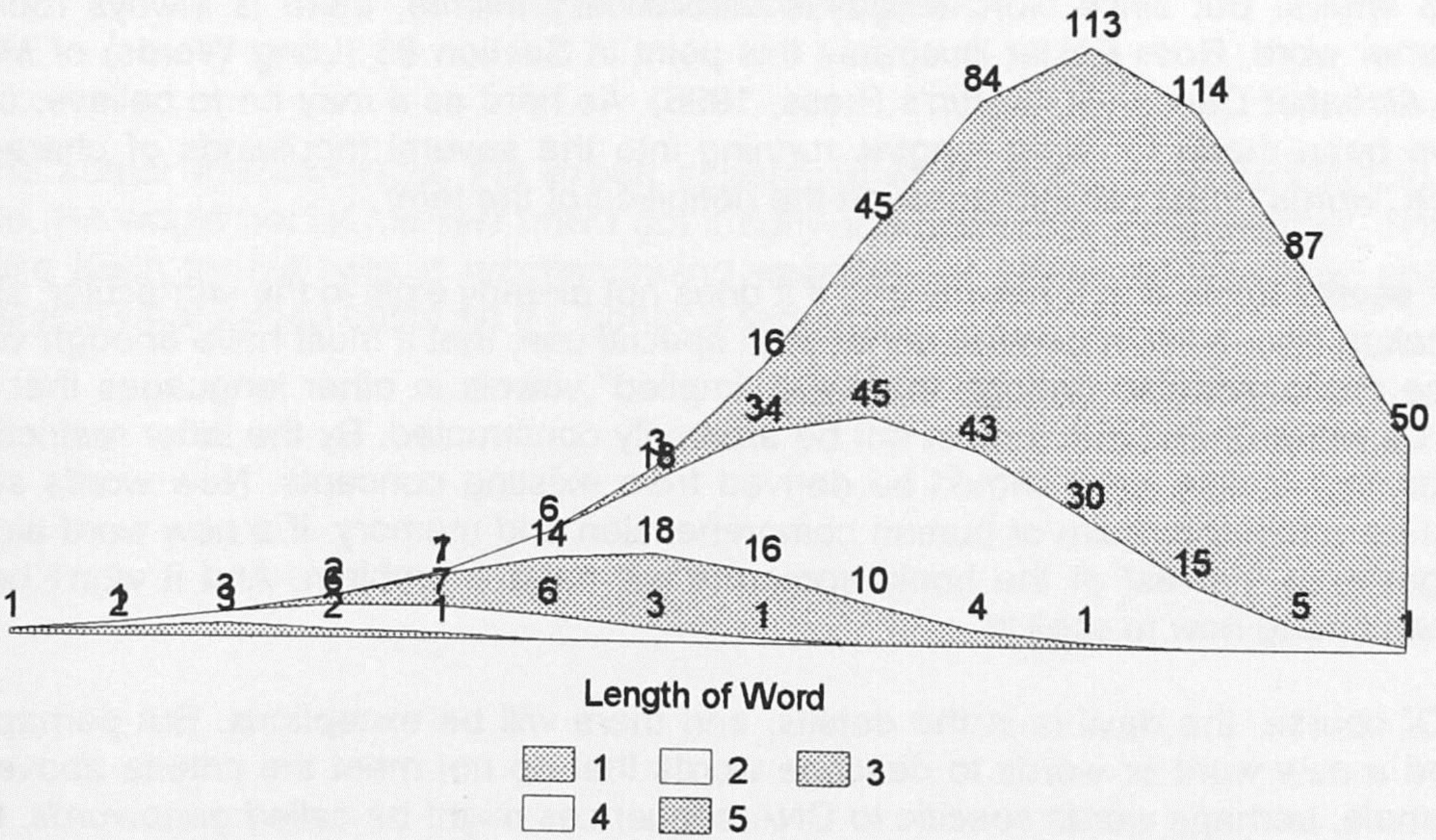
The sum of these four products is 3,606,720, the number of allowable 5-letter words. The complete set of vowel counts calculated for words of length 1 through 14 is given in the table and chart on the next page. The chart shows that as the length of a word increases from left to right, the number of vowels (indicated in the legend at the bottom of the chart) also increase, but that the shape of their distribution remains symmetrical.

The table below presents the maximum possible number of words of 1 through 14 letters, given that no word may have more than two consecutive vowels or consonants. The second column gives the percentage of allowable words of a given length, compared with the number if no restrictions are imposed (26 raised to a power equal to the word length). The sum of these numbers is 835,869,328,666,335,822. A number as large as this surely indicates that no language using a Roman alphabet of 26 letters is

Vowels >	1	2	3	4	5	6	7	8	9	10
1	1									
2	2	1								
3	3	3								
4	2	6	2							
5	1	7	7	1						
6		6	14	6						
7		3	18	18	3					
8		1	16	34	16	1				
9			10	45	45	10				
10			4	43	84	43	4			
11			1	30	113	113	30	1		
12				15	114	208	114	15		
13				5	87	285	285	87	5	
14				1	50	300	518	300	50	1

How Number of Vowels Change Shape

with increases in Word Length



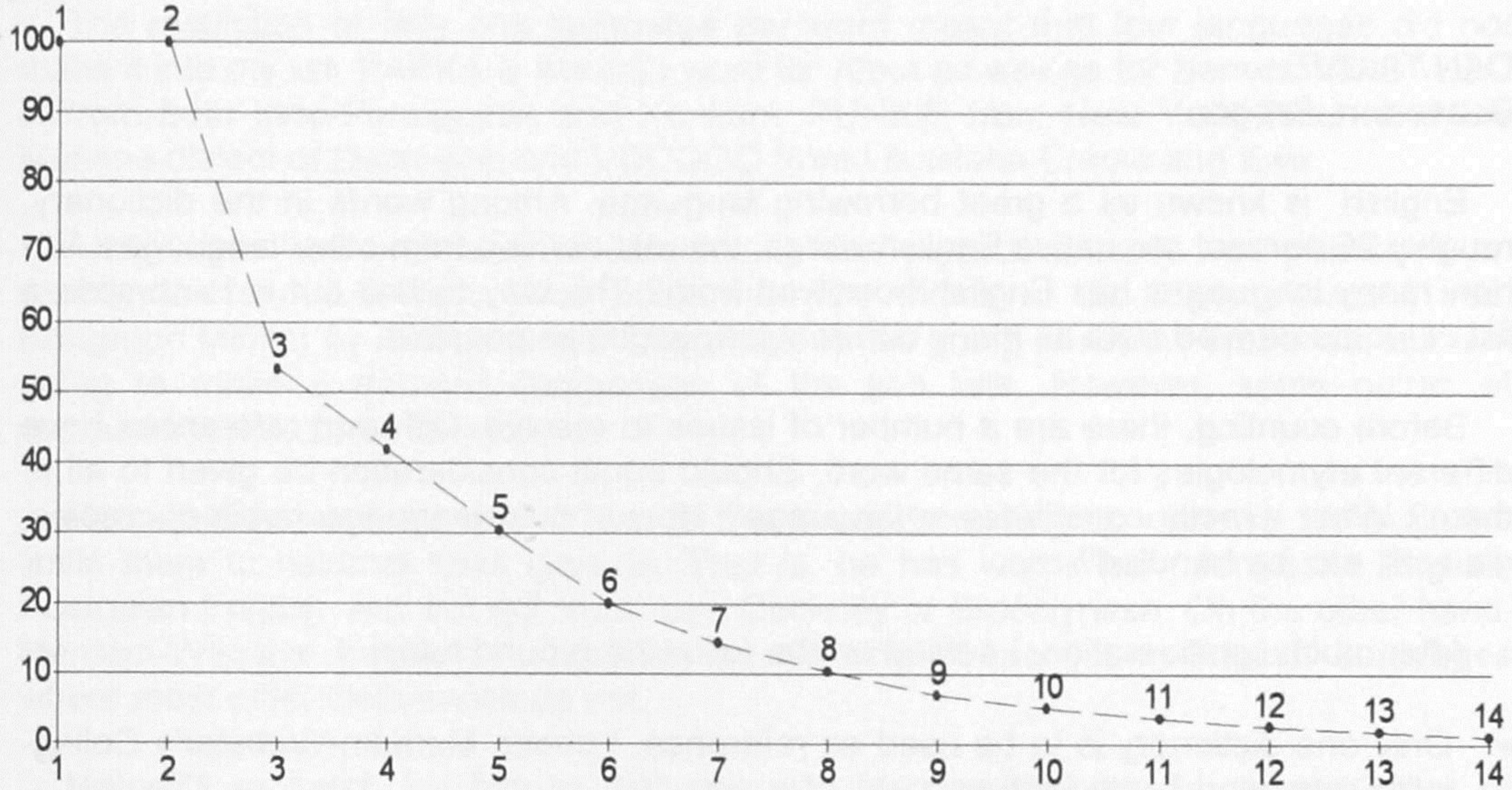
likely to run out of room for expansion anytime soon! (But if another letter or symbol were to be added, as in Norwegian, which has 29, the possibilities would greatly increase.)

1	100	26
2	100	676
3	53.3	9360
4	41.8	191040
5	30.4	3606720
6	20.0	61862400
7	14.5	1163635200
8	10.3	21427430400
9	7.08	384583680000
10	5.02	7092652032000
11	3.54	129759971328000
12	2.47	2359574323200000
13	1.74	43238275891200000
14	1.22	790134218588160000

The chart on the next page graphs the second column of the above table. The greatest opportunity for expanding the number of English words occurs at the middle levels. It seems reasonable that the easiest and shortest words have already been formed. For example, it may be difficult to find many more new words in the range of 1 to 3 letters. But since word length is theoretically infinite, there is always room for another word. Ross Eckler illustrates this point in Section 83 (Long Words) of *Making the Alphabet Dance* (St. Martin's Press, 1996). As hard as it may be to believe, claims have been made for word lengths running into the several thousands of characters! Such "words" must call into question the definition of the term.

It seems to me that a new "word", if it does not already exist in the vernacular, should be taken from a list in general rather than special use, that it must have enough vowels to be pronounceable (though there are "implied" vowels in other languages that need not be written), and that it must not be arbitrarily constructed. By the latter restriction,. I mean that a new word should be derived from existing concepts. New words should also be within the realm of human comprehension and memory. If a new word alone is longer than the rest of the book, someone will have a problem. And it won't be just remembering how to spell it!

Of course, the devil is in the details, and there will be exceptions. But perhaps we need a new word or words to describe words that do not meet the criteria above. For example, perhaps words specific to DNA sequences might be called *genowords*, those specific to chemisty, *chemowords*, those to medicine, *medwords*, and so on, as the demand may arise. In any case, these titanic words certainly lengthen the rightmost tail of the distribution of word lengths. Fortunately, as is pointed out on page 252 of *Making the Alphabet Dance*, the number of words at extreme word lengths tends to diminish by half with the addition of each new letter.



The author acknowledges the expert assistance of Ross Eckler in completing this article. He would not let me rest until I got it right--the mark of a good teacher. Thanks to Mike Keith for his help in recommending websites for dictionary wordlists, and for identifying what turned out to be a Fibonacci sequence.