The
PSYCHOLOGICAL RECORD

Vol. I NOVEMBER, 1937 No. 24

ON THE NUMBER OF WORDS OF ANY GIVEN FREQUENCY OF USE

EDWARD L. THORNDIKE

The Principia Press
Bloomington, Ind.

Price of this number, 20 cents

ON THE NUMBER OF WORDS OF ANY GIVEN FREQUENCY OF USE

EDWARD L. THORNDIKE
Institute of Educational Research, Teachers College, Columbia University

Zipf ['35, p. 39f and '37, p. 239f] has suggested that the well-known but unmeasured facts that there are very few very frequently used words, and increasingly larger numbers of less and less frequently used words, at least within certain limits, can be put in order to a first approximation by the equation ab^2=a constant for the given field of speech or writing, in which b=any number of occurrences and a=the number of words occurring that number of times. He has used the data of the Eldridge newspaper count of 43,989 running words as evidence in the case of English and samples of that order of magnitude in four other languages. He points out that this relation does not hold for very common words such as the articles, prepositions, conjunctions and pronouns, but reports a fairly close fit after a certain point and especially for rare words. He concludes, partly on the basis of these facts, that "The high degree of orderliness of the distribution of words in the stream of speech points unmistakably to a tendency to maintain an equilibrium in the stream of speech between the frequency on the one hand and what may tentatively be termed variety on the other." He adds, "Perhaps the most interesting feature of this high degree of orderliness in the distribution of words in the stream of speech is this: we select and arrange our words according to their meanings with little or no conscious reference to the relative frequency of occurrence of those words in the stream of speech, yet we find that words thus selected and arranged have a frequency distribution of great orderliness which for a large portion of the curve seems to be constant for language in general."

Skinner ['36, p. 86f and '37, p. 71f] has found a uniform relation of the same sort for the sixty or so words most commonly aroused by the obscure vowel sounds presented in his summator experiments, and also for the hundred or so words most commonly aroused by a stimulus word in the Kent-Rosanoff experiments.

I have recently completed a count of about 4½ million running words from 120 books taken with few exceptions from those recommended by Terman and Lima for supplementary reading by children in grades 3 to 8. In this count every derivative in s, ed, ing, n, ly, etc. (including 's, though the desirability of this may be questioned) was counted separately and all names of persons and places were included; but 2,500 of the commonest "lexical units" or names of persons or places in English (items 1 to 2,500 of the Thorndike 20,000) were not included, nor were most of the words formed from them by adding s, ed, ing, er, est and ly. This count satisfies Dr. Zipf's criteria except for the omission of these words' derivatives and the minor particular that words pronounced differently but spelled alike were not kept separate.* The effect of these omissions will be considered after the facts as they stand are reported. This count has also the merit of including many very rarely used words, though probably not so many as would an equally large sample from books intended for adults.

I have computed the values of ab^2 for words occurring from 1 to 44 times (i. e., from 1 in 4½ million to 1 in 100,000 running words) from three independent samples from this count. The first includes all words from cast-iron to Dyves. The second includes all words beginning in E, F, G and H. The third includes all words beginning in J, K, L and M to "mussel". They do not differ in respect to the fundamental issue, and I therefore combine them.

The value of ab^2 rises rapidly from the rarest words until words are reached which occur about ten times. Its value where b=10 is about five times its value for b=1. It continues to rise thereafter, but more slowly. The values for ab^2 are 7,519, 10,524, 13,932, 16,944, 19,875, 23,184, 24,451, 27,648, and averages of approximately 32,500, 35,500, 34,000, 40,500, 46,000, 41,000, 56,000, 56,500 and 55,500 for successive groups of four thereafter.

The relative magnitudes of these numbers would be altered only inappreciably by proper treatment of words pronounced differently but spelled alike. The effect of including the various forms of the 2,500 commonest words (by the Thorndike count of 1921) would be to make ab^2 still lower relatively for the words occurring only once, twice, or three times and higher for those occurring more than 8 times. This could be assumed on general grounds, but I have verified the assumption by a special count of the various forms of words ranking

* This was done deliberately to economize time and effort since the counters' work would have been much longer and harder if they had been required to consider the sound as well as the appearance of words.

from 2,501 to 3,000 in the Thorndike 20,000 list. What is true of these will be more emphatically true of words 1-2500. The effect of excluding proper names would also be to make the increase in ab^2 steeper.

I have computed the numbers of words that occur 45, 46, 47 . . . 120 times in the samples described, but their relations *inter se* and to the facts for the rarer words would be so much altered if the various forms of the commonest 2500 words were included that those not specially familiar with the frequencies of English words would probably be misled if I reported them here. By any reasonable allowances, ab^2 may be expected to increase through the range from frequencies of 1 per 4½ million to frequencies of 1 per hundred or less. $ab^{1.45}$ gives a fair fit for some distance from the rare end.

I have also computed ab^2 for words occurring from once to 80 times in the Anderson count ['17] of 361,184 running words in correspondence (except for words occurring 2, 3, or 4 times, for which the facts are not available). ab^2 rises here from 3217 for words occurring once to nearly 10,000 for words occurring ten times and to an average of 32,293 for words occurring 71 to 80 times.

Horn ['26] has reported frequency credit numbers for all non-capitalized words occurring 13 times or more in counts of 5,136,816 words found by him or others in correspondence. It is not possible to estimate actual frequencies accurately from these credit-numbers. I can, however, guarantee that, for words with credit numbers from 13 to 1,000, ab^2 is not a constant.

There is a peculiar danger in arguing from the words used and the frequency of their use in any sample of English speech or writing to the words that are used and their frequency in the total speech or writing of the person or persons sampled. Unless the sample is enormous in size (say 60 million words for juvenile books, or 200 million words for books for adults, 500,000 words for the speech of a peasant, or 40,000,000 words for the speech of 18-year-old American boys) it will seriously underestimate the number of different words by reporting 0 occurrences for a substantial number that would have occurred one or more times if the total usage had been observed. Even a peasant uses many words that he uses less than once a month.

There are literally tens of thousands of words which did not occur once in the 4½ million words of the juvenile books but which have occurred or will occur in other juvenile books (e. g., dairyman, dalliance, damnation, damson, dandruff, darnel, darner, dateless, davenport, and dayspring). It is certain that if Eldridge had counted all

the words in American newspapers in 1911 instead of a sample of about 45,000, he would have got nearer 60,000 different words than the 6,002 he did get, and nearer 29,000 words used only once than the 2,976 he did get. A more adequate sample will increase the number of words occurring only once or twice very greatly over that found in a small sample.

But it will *decrease* the percentage which the number of occurrences of each of them is of the total number counted. Suppose we increase a count a hundredfold. While the number of occurrences of certain rarely used words rises from 0 to 1 or 2, or from 1 to 3 or 6, or from 2 to 10 or 20, the number of occurrences of *and, the, in, on, of,* etc., rises to a hundred times its former amount. It will cause the value of ab^2 to rise more rapidly than it did in the smaller sample until a point is reached where the sample is adequate. The effect of using a juvenile count of 45 million in place of our count of 4½ million would be to make the values of ab^2 rise even faster, as we proceed from the rarest to the less rare, than in the data presented here.

What is so emphatically true of very rare words is true to a less degree of all words except those which are so common as to be surely included in even the small sample. For example, the number of words occurring from 1 to 60 times in the Eldridge count of 43,989 words, the Anderson count of 361,184 words and the Thorndike count of 4½ million words in juvenile books, and the percentage which each is of the total number of words counted, are shown in Table 1.

We may compare the ab percentages for words occurring once in Eldridge with the sum of those for words occurring 5, 6, 7, 8, 9, 10, 11 and 12 times in Anderson; the percentages for words occurring twice in Eldridge with the sum of those occurring 13, 14, 15, 16, 17, 18, 19, and 20 times in Anderson, and so on.

Even for words occurring 9 times in the Eldridge and 69 to 76 times in the Anderson the Eldridge still overestimates. These are such words as accept, among, article, auto, beginning, benefit, cash, circular, coal, commission, copy, couldn't, couple, cousin, demand, error, except, Feb., forget, held, hot, idea, issue, knew, low, March, May, o'clock, page, pleasant, probably, promptly, property, register, side, sign, snow, student, taking, trip, Tuesday, yard.

In the same way it can be shown that the Anderson overestimates the relative number of rare words in comparison to the 4½ million count.

The relation between a and b for all the speech or writing of any person or group of persons may then vary greatly according to the

size of the sample that is used. If ab^2 is a constant for words occurring from 1 to 1,000 times in samples of the size used by Zipf for the persons and groups used by Zipf, it almost certainly will not be that for the total usage of those persons and groups.

TABLE 1

The number of different words occurring once, twice, three times, etc., in the Eldridge newspaper count of 43,989 words, the Anderson correspondence count of 361,184 words, and the Thorndike count of 4½ million words in books for pupils in grades 3 to 8; also the percentage which each product (number of different words of one frequency × that frequency) is of the total number of occurrences in the samples.

	Eldridge		Anderson		Thorndike	
b occurrences	a Number of different words	Percent which a × b is of 43,989	a Number of different words	Percent which a × b is of 361,184	a Number of different words*	Percent which a × b is of 4,500,000
1	2,976	6.79	3,217	.89	18,547	.041
2	1,079	4.92	?	—	6,498	.029
3	516	3.52	?	—	3,824	.025½
4	294	2.67	?	—	2,613	.023
5	212	2.41	180	.25	1,964	.022
6	151	2.06	175	.29	1,591	.021
7	105	1.67	150	.29	1,233	.019
8	84	1.53	148	.33	1,067	.019
9	86	1.76	121	.30	818	.016
10	45	1.02	99	.27	795	.018
11	40	1.00	82	.25	647	.016
12	37	1.00	77	.25½	652	.017
13	25	.74	91	.33	511	.015
14	28	.89	68	.26	472	.015
15	26	.89	61	.25	346	.012
16	17	.62	69	.30½	338	.012
17–20	59	2.48	190	.97	1,099	.045
21–24	35	1.79	160	1.00	795	.040
25–28	31	1.87	152	1.12	647	.038
29–32	19	1.32	110	.93	437	.030
33–36	16	1.25	90	.86	467	.035
37–40	8	.70	79	.84	378	.032
41–44	13	1.25	62	.73	304	.028
45–48	9	.95	65	.84	220	.023
49–52	10	1.15	57	.80	178	.020
53–56	3	.37	40	.60	153	.019
57–60	3	.40	44	.71	128	.017
61–64	—	—	28	.48	98	.014
65–68	—	—	29	.54	108	.016
69–72	—	—	34	.66	106	.017
73–76	—	—	35	.72	93	.016
77–80	—	—	20	.43	70	.013

* Estimated reliably from the number found from "cast-iron" to "mussel" except words beginning with I.

If all the occurrences of all the different English words that were spoken or written or printed in the last year or decade or century had been counted, what would be the relation between the number of occurrences (from 1 or whatever the minimum number of occurrences was to whatever the maximum number was) and the number of different words having that number of occurrences. No adequate answer can be given, but certain limits can be set up. (1) The total number of different words for the last decade was not over 5 million, probably not much over a million. (This counts forms in s, ed, ing etc., separately, and includes names of persons and places, a catholic list of compound words and all provincialisms and slang, but not infantile variants nor such compound number names as "sixty-two thousand four fundred eighty-five.") (2) Let t equal the total number of different words, and T equal the total number of occurrences. Let l equal the minimum number of occurrences of any word that occurred at all. Let $l + k$ equal the number of occurrences of the word that occurred oftenest (almost certainly "the"). Then l will be a small number, almost certainly under 100 for a decade and possibly as low as 1; $l + k$ for a single year will be above 1,000,000,000,000 (and, of course, about ten times as large for a decade). There will be words occurring $l + 1$ times, $l + 2$ times, $l + 3$ times, etc., up to $l + 1{,}000$ times with few or no exceptions, but there will be more and more instances of 0 occurrences as we go from $l + 1{,}000$ to $l + 10{,}000$, and in the range of $l + 1{,}000{,}000{,}000$ and over the gaps will be numerous and wide. From l to $l + .01\ k$ the number of different words will be well over $.99\ t$ and the number of occurrences well under .01 T. From $l + k$ down to $l + .99\ k$ the number of different words will be well under $.00001\ t$ and the number of occurrences well over .05 T.

For a random sample of 100,000,000,000,000 spoken, written and printed words of 1925-1935, the magnitude of t for each equal division of $.01\ k$ along the scale from l to $l + k$ will fall very rapidly from something in the millions to something probably less than ten. The magnitude of T for each equal division will rise from something in the order of 10,000,000 to something in the order of 5,000,000,000,000.

The average values of ab^2 (or of a times any power of b) for any division of k will obviously be very different according as the gaps (where a is zero) are excluded entirely or are treated as zero values of ab^2. Either procedure may be justifiable according to our purpose. For example, in the Anderson count of 361,184 occurrences of 9,224 different words used in adult correspondence the fifteen commonest words are: I, and, the, to, you, of, in, your, for, we, is, that,

it, have, are. The 15 values of ab^2 for these words taken separately average about 55,000,000; the 8400 values of ab^2 for words occurring 12,000 times, 11,999, 11,008 times . . . 3600 times average about 100,000. Both statements of the fact are true. The occurrences of these 15 words out of 9224 include over a fourth of T (which is 361,184). But we have to go seventy hundredths of the way down from $l + k$ (which is here about 12,000) toward l to get these 15 words and this fraction of T. Most of the space from $l + \frac{2600}{12000}k$ to k is empty of words.

It is useful to compare both numbers with the 3,217 which is the value of ab^2 for words occurring once in the Anderson count, or with the 12,429 which is the average of ab^2 for words occurring 11, 12, 13 or 14 times, or with the 20,155 which is the average ab^2 for words occurring 21, 22, 23, or 24 times,* or with the 27,952 which is the average ab^2 for words occurring 41 to 44 times.

Using the second method, I estimate that for total English usage ab^2 for words from $l + .3k$ to $l + k$ is at least 50 times what it is for the words at l, $l + 1$, and $l + 2$ and $l + 3$. The ratio is probably much above 50 to 1.

It seems to me unlikely that the relation for such total usage, no matter how uniform it turned out to be, would be evidence of any uniform and ubiquitous tendency toward a certain equilibrium between frequency and variety. I should expect that it would be in some measure a statistical artifact, adding things which lost much of their instructiveness by being combined. The relation between a and b presumably varies not only with the size of the sample, but also with the natures of the persons who speak and write, their knowledge of English, their purposes in speaking and writing, their past experiences, their access to the speech and writing of others, and whatever else makes them and the language what they and it are. The relation may then be expected to vary at certain points and perhaps at all points among different persons and different groups and between technical writing and popular writing, writing for adults and writing for juveniles, writing stories and writing letters, etc., etc. I fear, therefore, that before we can use the relation between frequency of use and the number of words having each frequency safely in either linquistics

* To be reported by Anderson they must have occurred (except for the special category of single occurrences) in at least 3 out of his 6 varieties of correspondence. The values of ab^2 may therefore be a little low for a complete count of the 11-14 words, and possibly even for the 21-24 words.

or psychology we will have to determine it in many instructive cases, increasing the sample in each case until the relation becomes stable.

There surely are forces acting to cause variety. The law of effect makes speakers and writers tend to abbreviate a statement originally made with many common words. They may abbreviate it in all sorts of ways, one of which is to coin a new word, and this has become an orthodox procedure for new colors, fabrics, trades, machines, processes, etc., in spite of occasional remonstrances from hearers and readers. Variety in the world thus causes variety in vocabulary. Something fundamental and possibly related to the refractory phase of brain activity seems to make us averse to using a word too soon after we have spoken or written it. There is also a tendency for both speakers and hearers to enjoy certain varieties more than certain monotonies in speech as elsewhere. There is, for literary men, a peculiar zest in inventing linguistic novelties.

There surely are forces acting to cause frequent use of the same word. Certain persons, things, events, and relations occur often in thought and communication. Other things being equal, the use of one symbol for one person, thing, event or relation saves much time and effort for the speaker-writer, and still more for the hearer-reader.

That there is a force wider and deper than any of these acting to produce equilibrium for equilibrium's sake, I do not believe. Nor does Dr. Zipf, I think, though his words may be so interpreted.

REFERENCES

Anderson, N. N., '17—Determination of a Spelling Vocabulary Based upon Written Correspondence. University of Iowa Studies in Education. Vol. II, No. 1.

Horn, E., '26—A Basic Writing Vocabulary. University of Iowa Monographs in Education, First Series, No. 4.

Skinner, B. F., '36—The Verbal Summator and a Method for the Study of Latent Speech. Journal of Psychology, Vol. 2, pp. 71-107.

Skinner, B. F., '37—The Distribution of Associated Words. Psychological Record, Vol. 1, No. 6.

Zipf, G. K., '35—The Psycho-Biology of Language.

Zipf, G. K., '37—Observations of the Possible Effect of Mental Age upon the Frequency-distribution of Words from the viewpoint of Dynamic Philology. Journal of Psychology, Vol. 4, pp. 239-244.