

## *How Does "THE" Grab You?*

ARLAND R. MEADE

NOTHING CAN BE MORE OBVIOUS than that the failure to write effective letters leads to business failure of any concern that lives by the effectiveness of its sales letters. The life of the Extension Service is not so vulnerable, but, at best, tax money would be wasted on a "failure type" direct mail circular or letter.

So it occurred to me to find out what "miracle words" are used by the successful practitioner of direct mail letter writing — and the companies they succeed with. Does every message start out with an intriguing, a catchy, a dramatic, a patriotic, or a personal word? Are paragraphs constructed so that they would each start off with a word that is eye catching, significant, heart-throbbing, prestige-suggesting, motivating, ego-inflating, keep-up-with-the-jones-ing, or whatever moves the recipient to action?

So during some midnight hours of the past few years I have been tabulating words of seemingly successful sales letters.

I tabulated first words of every paragraph of every letter addressed to me at my home or to me at the office — from organizations which were either (1) soliciting money for a cause, or (2) trying to sell me a product or a service. These were the basic parameters.

Both types were "selling" letters, inasmuch as the proponents of the most worthy of causes were still selling me on the idea of parting with money for their support; and the others were obviously asking me to part with money for a product or service.

I did not include local solicitations — only those from organizations which were widespread, surely with professional letter writers, and who seemingly were in a sink-or-swim situation depending on the effectiveness of their direct mail.

Word counting is tedious and terribly time consuming. However, there is much linguistic precedent for statistical analyses of

language through counts of words, phonemes, or other components. Fortunately when Morse determined the dot-dash symbols for letters in English, he did a great deal of such investigation. That is why the briefest signal, one dot, is for the letter that appears most frequently in English (e) and the longest signals are for the letters that appear least frequently (q, j, x, z).

In this paper I won't delve deeply into citations or references. I will, however, mention a news release this fall which revealed that a dictionary publisher had spent about three million dollars in counting words relative to a new edition. We can assume there was more to it than simple counting, and still feel a bit of awe for this statistical approach.

So, foolhardy soul that I was, I started — with occasional help from Colette Beauchesne, secretary — to list and tally initial words of every one of the 7,522 paragraphs in my stack of 522 letters.

When the task was completed, I found that the magnificent word "THE" was on top. It was the initial word of 6.9 per cent of all the paragraphs (520 times it appeared of a maximum possible of 7,522 appearances).

Runner up was the word "YOU" which initiated 5.3 per cent of all paragraphs (398 appearances).

### *Two Words Start One in Eight*

Of the approximately half million words in the English language, only two of them were thus used to initiate 12.2 per cent of all paragraphs in our sample—a sample of writings that we have accepted as critical to the writers and by and large effective on the receivers.

This is so far from random chance that it staggers the imagination.

Well, what other "impactful" words initiated these paragraphs?

How about "AND." That was third with 3.6 per cent of the total (276) appearances.

Followed by "IF" with 2.9 per cent (219) of the appearances and "WE" with 2.8 per cent (208) of the appearances.

Now what would be a reasonable vocabulary of the professionals who wrote these letters? Studies indicate that the vocabulary of a typical American college graduate is likely to be upward of 20,000.

From whatever it was, the writers led off one-fifth (21.5 per cent) of all their paragraphs with just five words: THE, YOU, AND, IF, and WE. Oh wow!

Well, surely some "pulling" words must soon come into the picture: maybe, please, send, love, save, immediately. . . .

Nope! The next five words come up: THIS, IN, BUT, TO, A. These five initiated 798 paragraphs—10.6 per cent of the total.

So the 10 words most frequently used to begin paragraphs have initiated 2,419 of the 7,522 paragraphs: 32.1 per cent.

Let's try again. The third most frequent five were: IT, I, YOUR, AS, HERE. This quintet accounted for 598 initial uses: 7.9 per cent.

Another round of five brings up FOR, HOW, SO, WHEN, NOW. These account for 375 paragraphs (5.0 per cent). The top 20 haven't quite accounted for half of all paragraphs, (45 per cent) so let's see how far we must go to reach that point.

Words in frequency sequence follow: YES, WHY, JUST, WITH, NO (304 appearances).

These top 25 words now account for 3,696 paragraphs. Only 61 more paragraphs will equal half of the sample. So we move on to "ALL" with 59 uses and we are only two paragraphs short of one-half of the entire sample.

It took **only 26 different** words to initiate 50 per cent of all paragraphs in our sample of 7,522 paragraphs. Oh wow! again.

Do you see where to place your bets?

### *Shock Value Important*

Information theorists tell us that the **expected** carries little or no meaning. If, day after day, a certain man always says "Good morning, how are you" when he meets me, I have to stretch to find any information there—other than that he's still alive. But the morning he just growls and looks the other way I know that something is changed; he has signaled some information.

Now the initial words of paragraphs aren't quite so predictable as some people's "Good morning, how are you." Still, when one word "the" initiates seven per cent of all paragraphs, and we are going to get either "the" or "you" in 12 per cent, and when of the multithousands of English words only 26 are going to greet

us at the beginning of half of all paragraphs, one sees a great deal of redundancy there and very little information. (I'm not challenging the valuable functions of redundancy, but that's another story.)

### *Does Language Demand Article?*

Should we assume that there's something about the syntax of our language that forces us to this selection of articles, prepositions, and conjunctions that so typically start our paragraphs? And is this true of all sentences in the same type of writing? Of all writing?

This I'm not able to say. The dictionary study did prove that the word "the" was the most used in all the language. Yet the definite article does not even exist in Russian.

I have many other statistical facts from my study: length of paragraph, length of letter, use of colors and ornaments, etc.

Perhaps I'll report those another day, but for now can we derive a communication fact or two from what I've found in tallying words? For example, if the words so generally used to start sentences are not carriers of information, should writers of sales letters force themselves to use words that are? Would "livelier" words better attract the reader's attention? Or would we somehow burden the reader if we began with biff, bam, and wow instead of the, and, or but?

There clearly is prediction value—"heuristic" is the term the communications scientists use—in what I've found.

No doubt I could win some bets: "I'll bet you, Mr. extension direct mail letter writer, that I can tell you which words initiated half of your last hundred paragraphs—and not even know what the subject matter is."

To test this, I predicted the frequency order of words in a collection of letters I'd thrown in a file, and not even read. These all arrived after I closed my big study sample. I predicted in this order: the, you, and, if, we, this, in, but, to, a. My top 10.

Then I tallied the paragraph-initiating words for all these letters and here is what I found.

For comparison I've listed the predicted first 10 words in sequence and immediately following have put the sequential position of the same word in the test sample of 25 letters (con-

taining 310 paragraphs). Of the top five in the study, four were in the top five of the test sample and one stood sixth.

- (1) the (1)
- (2) you (3)
- (3) and (6)
- (4) if (4)
- (5) we (2)

To continue:

- (6) this (5)
- (7) in (7)
- (8) but (11)
- (9) to (not in top 20)
- (10) a (not in top 20)

Of the top 10 in the study, seven were in the top 10 of the sample and three were in the same positions. Four others differed from one to three positions; one placed 11th; and—a surprise—two did not make it to the top 20 of the test sample.

With a prediction confirmation of seven out of 10, I probably won the bet. I say “probably,” because with a sample of 310 paragraphs my chances are much better than with the 100 of the “bet.” That’s too small a sample without some odds. (Would some wager-minded extension editor offer such a challenge to a few extension specialists and tell me the results?)

Anyway, I continued the comparison through the top 20 words of the study. Beyond these there is so much scatter in the 25-letter test sample that comparisons would be of little value.

- (11) it (9)
- (12) I (12)
- (13) your (17)
- (14) as (8)
- (15) here (not in top 20)

Of the top 15 of the study, 11 were in the top 15 of the test sample; one placed 17th; and three were not in the top 20 of the test sample.

- (16) for (18)
- (17) how (14)
- (18) so (15)
- (19) when (10)
- (20) now (not in top 20)

Of the top 20 in the study, all but four were in the top 20 of the test sample.

For predictive value, my study is clearly high when applied to a separate sample of direct mail letters of the same parameters (outlined in the early part of this article). Except for these parameters, selection was random.

Word sequences of the study and the sample were in close approximation, with the exception of the words "to" and "a" which placed ninth and 10th in the study but failed to make the top 20 in the sample. "To" did not show up as initiating ANY paragraph in the sample, and "a" would have needed two more appearances to have tied for 20th position.

One can question whether these commonly used initial words are of any use. Students of information theory state that without surprise value there is no information. According to this study, the predictability of initial words of paragraphs is so great that surprise value must be slight.

### *Other Word Counts*

There have been many word counts, some done with many workers, such as the Irving Lorge semantic counts financed by the Rockefeller Foundation and others. The other counts that I'm familiar with counted running words, not initial words in sentences.

Is there a difference?

I compared my most-used 20 words with the top 20 of five other studies—by Dewey, Fraprie, Fossum, Voelker, and Black and Ausherman. There was more similarity among those studies than between mine and theirs.

For example, of my top words:

- (1) Six do not appear in any of the other five lists.
- (2) Seven appear in all the other lists also.
- (3) Nine words appear in three or more of the other five lists that do not appear at all in my top 20.

Note that the nine words common to other lists based on running language but absent in my top 20 initial words are: of, that, is, be, was, are, he, have, they.

Also that the six words in my top 20 which appear in none of the other five top 20's are: if, your, here, how, so, and when.

Does simple inspection suggest something here? That, for example, successful direct mail paragraphs (sentences) do not begin with forms of the verb "to be," are, was, but they often begin with if, your, here, how, and when.

To me, contrasts between word order in my initial paragraph words and a sample of 288,152 words uttered in 607 in-college speech recordings by Black and Ausherman are intriguing.

Although "the" placed first in both, and "and," "it," and "in" were within two steps of the same order, what happened to "if" that was fourth in mine and 48th in Black's study. And why would writers of direct mail letters start sentences with "here" in 15th place when it stood 121st in the running words of speeches. Likewise with my seventh place "but" which was down to 26th in the speech list; "how" which I had at 17th but which was 111th on the other.

I submit that these striking differences are not just because different samples were counted, but that different kinds of samples were tallied; and that word frequencies in discourse are not the same as in the initial words in written sentences.

Repetitiousness of our words is well documented by Irving Lorge, by Black and Ausherman, and others. Black and Ausherman found that of 288,152 word symbols uttered, approximately 15,000 of them were "the" and only 42 different words accounted for 131,758 word-symbols uttered.

Furthermore the top 371 words accounted for 224,645 of the total of 288,152 word-symbols recorded.

Now back to my study. Are the meagre selection of initial words in sentences just a syntactical habit? Or could they be a syntactical must? (We cannot, of course, stray far from the syntax of our language.) Or could such word choices be a colossal "bomb" never discovered by their writers or readers? Perhaps the surprise, hence the information value, of the paragraphs comes in their second words not their first. At the moment I decline to look.

But, as my secretary took a second look at paragraph beginnings of this article, she reminds me that of 63 paragraphs, **not one** begins with **any** of the five words which began one-fifth of the 7,522 paragraphs of my basic study.

Could I be doing something wrong?

## References

- (1) BLACK, JOHN W. AND MARIAN AUSERMAN. 1955. *The Vocabulary of College Students in Classroom Speeches*. Columbus: Bureau of Educational Research, Ohio State University.
- (2) LORGE, IRVING AND EDWARD L. THORNDIKE. 1938. *A Semantic Count of English Words*. New York. The Institute of Educational Research, Teachers College, Columbia University.
- (3) LORGE, IRVING. 1937. "The English Semantic Count," *Teachers College Record*, 39, 65-77 (October).
- (4) LORGE, IRVING. 1949. *The Semantic Count of the 570 Commonest English Words*. New York: Bureau of Publications, Teachers College, Columbia University.
- (5) FOSSUM, ERNEST C. 1944. "An Analysis of the Dynamic Vocabulary of Junior College Students," *Speech Monographs*, Vol. XL, pp 88-96.
- (6) FRAPERIE, FRANK R. 1950. "The twenty commonest English words . . . from a count of 242,432 words of English text taken from fifteen English authors and many newspapers." *World Almanac and Book of Facts for 1950*, edited by Harvey Hansen. New York: World-Telegram, p. 543.
- (7) THORNDIKE, E. L., AND IRVING LORGE. 1944. *The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications, Teachers College, Columbia University.
- (8) SHANNON, C. E. 1951. "Prediction and Entropy of Printed English," *Bell System Technical Journal*, III (January) pp. 50-64.
- (9) VOELKER, CHARLES H. 1942. "The One-Thousand Most Frequent Spoken Words," *Quarterly Journal of Speech*, XXVIII, pp. 189-97.
- (10) DALE, EDGAR. 1949. "Bibliography of Vocabulary Studies," Columbus, Ohio: Bureau of Educational Research, Ohio State University, v + 101 pp.
- (11) DEWEY, GODFREY. 1923. *The Relative Frequency of English Speech Sounds*. Cambridge, Massachusetts: Harvard University Press.
- (12) FLESCHE, R. F. 1948. "A New Readability Yardstick." *Journal of Applied Psychology* 32:221-33.
- (13) FUNKHOUSER, G. RAY AND NATHAN MACCOBY. 1971. "Communicating Specialized Science Information to a Lay Audience," *The Journal of Communication* Vol. 21, p. 58-71. Lawrence, Kansas.
- (14) WIENER, NORBERT. 1967. *The Human Use of Human Beings*. New York: Avon Books.
- (15) CHERRY, COLLIN. 1967. *On Human Communications*, 2nd edition. London England and Cambridge, Mass.: The M.I.T. Press.