CHANCE REMARKS

BY J. J. COUPLING

Remember "Fifty Million Monkeys"? There was a basic idea in that, and this pure fact article recounts an actual analysis of the problem of a "semantic selector"—and how it might work! A truly fascinating bit of communication research it is, too!

There has been a lot in the papers about Cybernetics and the theory of communications recently, but as usual, Astounding SCIENCE FICTION was ahead of all others—back at least as far as 1943 and Raymond F. Jones' story "Fifty Million Monkeys." You may remember the "semantic analyzer" which selected meaningful words from random letters. Now something very like that has appeared in the most respectable sort of print.

It was a long and indirect route if any which led from Jones' story to an authoritative publication, "A Mathematical Theory of Communication," by Dr. C. E. Shannon. Dr. Shannon, who is now at the Bell Telephone Laboratories, is a product of the Massachusetts Institute of Technology and of the Institute for Advanced Study at Princeton. His accomplishments include the application of Boolean Algebra—a symbolic logic—to the problem of telephone switching. His new work, which won him the Morris Liebman Memorial Prize of the Institute of Radio Engineers, is a fine exercise in multidimensional geometry and probability theory.

The whole field that Dr. Shannon covers is important, but it is so broad as to make a simple explanation exceedingly difficult. Too, much of the material, while of great technical importance, may seem to have little interest for any one but an expert. But part I section 2 of the paper, "The Discrete Source of Information," deals with something which seems right up the science-fiction alley. That is, the statistical structure of written English.

The American Telephone and Telegraph Company has been interested in transmitting English in one form or another for many years. I suppose it was inevitable that sooner or later someone should examine the material which they handle. But, when one first reads Dr. Shannon's words it seems that a language, and written English to be specific, means something quite different to him from what it does in common usage. This is natural. One believes that a communications company cannot afford much interest in the sense of the

material which it transmits. A correct reproduction is required. Whether or not the words sent are worth sending, or even, whether they make sense, cannot be of immediate concern. Because of this, one might wonder whether there can be any general interest in even this part of "A Mathematical Theory of Communication." Such an interest may, I think, develop in the telling.

First, however, one must know why the telephone company is interested in English, and what it is to them. Although this involves very long-range considerations, it can be made clear by a simple example. If you have ever watched a man trying to write a Christmas telegram, you may feel that he profited in the end by the telegraph company's neat arrangement. A few numbered messages express tritely about all that most people can say on the occasion. The cost is that of sending a short number instead of many words. The operator at the far end looks the message up by number in a little catalog, transcribes it, and the original message thus reaches the addressee without having been sent word by word at all.

It might seem absurd to apply such a principle to English text in general, but, in theory at least, this is merely more difficult. There is only a finite number of messages, that is, of combinations of symbols, which could be typed on a sheet of paper. Thus, we might merely type out all such messages, number them, and send the numbers instead of the messages. It turns out that this would result in no saving, because there are so many of such "messages". Even if we used double-line spacing and restricted ourselves to capital letters, spaces, commas and periods, the number would be about one followed by twenty-five hundred zeros.

Obviously, the suggested numbering of all of these messages is impossible. What is worse from a mathematical point of view is that it would be grossly inefficient. Many among the possible combinations of symbols would never be sent, at least, not in this country. More are in German and Italian and Swedish and French and in other languages than are in English. Some disclose the government's most closely guarded secrets—but which? Some are subversive. Many others would be censored. But, most of them are completely unpronounceable. They are combinations of symbols which make no sense at all.

It is, in fact, immediately clear that English text is in some way set aside from mere combinations of letters. It is set aside in a way which is important to one who is trying to transmit information. Subjectively, one easily tells whether text is English or whether it is not. There is, however, an objective distinction.

Dr. Shannon describes this by saying that written English is *redundant*. That is, more symbols are used than are needed to convey the information. Now, what is im-

portant is that the excess symbols are introduced according to certain rules, which a mathematician calls statistical laws or probabilities. For instance, *q* is always followed by *u*. Most of the rules are not this simple. However, a writer in mid-word or in mid-sentence does not—ordinarily—exercise complete freedom of choice in setting down the next letter or word. The choice of some letters or words is completely ruled out, or, such choices have zero probability. Among the letters or words allowed, some are more probable than others. For instance, because of our unconscious knowledge of such statistical rules, or of certain specific instances of them, we can correctly reconstruct most text—the reader can easily verify this—even if many of the individual letters have been erased or struck out. When we cannot, as in the case of some passages from Gertrude Stein or James Joyce, it is because there is something objectively un-English about the original text. That is, the text doesn't follow the rules. Granted conventional English text as a source from which to draw statistical rules, it is theoretically possible to tell English text from un-English sequences of symbols by applying statistical tests.

Dr. Shannon's work has led him to believe that English is about fifty percent redundant, that is, that one has about half as much freedom of choice as if symbols could be chosen with complete freedom. This is no trivial observation, for it has all sorts of implications. For instance, this degree of redundancy makes it just possible to construct crossword puzzles. If there were no redundancy, any arbitrary combination of letters would be a word. Thus, any set of letters could be read up or down as well as crosswise, and backward, too. There would be no puzzle to crosswords. If the redundancy were much greater than fifty percent, there would be so little freedom of choice in the sequence of letters that it would be impossible to achieve a crossword pattern. The degree of redundancy of English allows the construction of crossword patterns with some difficulty, and makes the language almost ideal for crossword puzzles. Conceivably, crossword puzzles could not succeed in some countries because the structure of the language would make them virtually impossible.

It is this same redundancy which potentially allows a saving in transmitting English text. It is cheap to telegraph a Christmas greeting because the message sent is chosen from among a few possibilities. Because there are rules relating the combinations of symbols, these restrict the number of English messages of a given length. The selection and numbering of page-long English messages suggests itself because these form a—comparatively—small set among possible combinations of letters, spaces, periods and commas, and the nonredundant num-

bers describing this small set would suffice for all sane Americans. But how many un-English messages are there to be discarded, and what can one do about it? This, essentially, is the language problem as it appears to a mathematician.

To the mathematician a language is a "stochastic—i.e., a statistical—process which generates a discrete sequence of symbols from a finite set." These symbols are the letters of the language, together with punctuation and spaces, if these occur. The stochastic process chooses these symbols in accordance with certain probabilities which involve the sequence of symbols already chosen. Thus, if part of a word or a sentence has been written down, the probability, as evaluated from ordinary English text, that the next letter will be *a* may be very high, while the probability that the next letter will be *e* may be very low, and these probabilities will depend on the preceding letters and the order in which they occur—that is, on what has already been written down.

If the statistics of the language were completely known, it would be possible—again in theory—to evaluate exactly the saving which could be made in transmitting English text. It would also be possible to do other things of which we shall have a hint later. Of course, a knowledge of the whole statistical structure of a language is an unattainable ideal, but one need not for this reason forgo all knowledge. Indeed, Dr. Shannon has done a little preliminary exploration himself in a surprisingly simple and a rather interesting manner.

We already know from work on cryptography, and can obtain from other sources, a small part of the statistical laws of English text. Now, suppose we choose symbols—letters—by a chance process incorporating the rules which we know and see how nearly the result resembles English. This will give us some clue as to the relative importance of the part of the statistical rules which we know and employ, and the unknown part of the statistics of the language.

Dr. Shannon gives first an exceedingly simple example:

"XFMOL, RXKHRJFFJUM ZLPWCFWKCYJ FFJEYVKCQ-SGXYD QPAAMKBZAACIBZL-HJQD"

Here the letters and spaces are succesively drawn at random with equal probabilities for all symbols. Here, for instance, *x* and *z* are as common as *e* and *a*, which they certainly are not in English text. The combinations are un-English, unpronounceable and uninteresting. Mathematically, we say that the statistics are incorrect.

"OCRO HLI NMIELWIS EU LL NBNESBYA TH EEI ALH-ENTTPA OOBTTVA NAH BRL"

Here letters were chosen, still independently, but with regard for their probabilities in English. If they were chosen from a hat, there would be more *e*'s in the hat than *z*'s, for instance. There is still, how-

ever, no rule connecting pairs of letters; there is no rule saying that *u* is the only letter which has any probability at all of following *q*. Still, some of the statistics of English have been taken into account, and the result is surprisingly more like English than the first sample. The letters do occasionally form word-like combinations. Although it is not in the dictionary, OCRO is pronounceable. It is interesting to think of OCRO as a nonsense word, and to wonder who invented it. It was really begotten, although through a human agency, by an undistinguished copy of a book of random numbers. A machine following the rules could as easily have arrived at the combination.

"ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE"

This has some intriguing features. Here more of the statistics of English were observed, and each letter or space was chosen in accordance with its probability of following that preceding. If *q* had occurred in the example, it could have been followed only by *u*. This has resulted in combinations of symbols which strongly resemble English. ARE is a real word. INCTORE and ILONASIVE are not words, but they are very wordlike. DEAMY almost suggests meaning, and is perhaps worthy of remembrance. One begins to wonder if Dr. Shannon's work has some literary significance. Can senseless statistics perhaps aid to our vocabulary?

"IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE"

To an unprejudiced reader, this is not only largely pronounceable, but it sounds like talk, and English talk more than anything else—it contains seven Engilsh words. Or, if you wish, the passage sounds like double talk. It is perhaps difficult to believe that in constructing this passage no conscious effort was made to make up English-like combinations. The procedure of construction was, however, purely automatic; each letter was chosen in accordance with the probability of its following the ordered pair of letters preceding it.

It must be understood that, because of the increasingly elaborate statistics involved, these passages were increasingly difficult to construct. A complicated machine could do more, but without one it seemed impractical to go further, and to base a letter on a preceding ordered triplet of letters. We can, however, see where the process would lead. If we added letters according to rules involving three, four, five and more preceding letters, we would gradually rule out as of zero probability all combinations which do not appear as words in the dictionary. We might as well,

indeed, use words as our basis of choice, and Dr. Shannon has tried this, too. As a first example he chose words merely on the basis of their probability of appearing in English text:

"REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME CAN DIFFERENT NATURAL TO FURNISHES THE LINE MESSAGE HAD BE THESE"

This seems rather a retrogression. The statistics are unduly simple, for they provide no connection between words. With a great deal of effort, Dr. Shannon was able to provide such a connection, however. In obtaining the following passage, a pair of words was chosen at random in a novel. The novel was then read through until the second of these words was encountered again, and the word following it was inserted. Then that new word was sought out in a new context, and the word following it there was added, and so on. This laborious process evoked:

"THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED"

Here we have merely an example of words chosen randomly according to certain statistics. We may, however, have a strange feeling that we have seen something like this before. Certain passages in "Ulysses" and "Finnegan's Wake" are scarcely more intelligible. Despite an apparent lack of connection, the passage has some subjective interest. I have a sympathetic concern for the predicament of the English writer. I would like to ask the author more about him. Unfortunately, there is no author to ask. I shall hear no more unless, perhaps, chance should answer my questions. One wonders if Dr. Shannon's work has philosophical implications.

Dr. Shannon stops at this point. The idea of pursuing the matter further is, however, tempting. By taking into account more and more statistics in the choice of letters, all letter combinations but English words could be ruled out. Can we, by the use of a more elaborate statistical choice of words, rule out all word combinations which don't make sense? At least, we could construct something better organized than the last example.

Suppose, for instance, that a word were chosen in accordance with its probability of following the preceding three words. Although this might seem unduly difficult, a trick will overcome all obstacles. English and its statistics reside in the human brain and they can be tapped at the source. One has only to show a list of the latest three words of a passage to a person unfamiliar with

those preceding and ask him to make up a sentence including these three words and to write down the word which, in that sentence, follows the three. The statistics linking four-word combinations are automatically evoked in this process. The word chosen *can* and *is likely to* follow the three. There is, however, a chance element in the choice. The word chosen is not *determined* by the preceding three words, for different people, or the same person at different times, would choose a different word.

In following this procedure we can also, without added difficulty, include punctuation and capitalization. This further lends naturalness to the result.

Starting with the words, "When the morning—" I obtained from twenty-one acquaintances:

"When the morning broke after an orgy of abandon he said her head shook quickly vertically aligned in a sequence of words follows what"

This begins well, and the eighteen words following the initial three have a clear meaning. Afterwards there is a wandering of the mind, as in some cases of schizophrenia*. But whose is the meaning, and whose mind wanders? We must admit that the meaning exists only in the minds of the readers. Each of the twenty-one writers knew only four words, and each thought of them in a different context. There was no "meaning" until someone read the completed passage. And there was no wandering of the mind, but only failure of such short-range statistics as were taken into account to hold the text together over many words. The words are connected to those immediately preceding them, but have no connection with those further ahead. I think, however, that it is the seeming sense of the passage and not its long-range incoherence that is astounding. And, it is a little disturbing to think that an elaborate machine, taking longer-range statistics into account, would have done still better. The passage seems to us to have meaning, and yet the true and only source of this quotation is a small part of the statistics of the English language—and chance.

Presumably, written English is coherent over long stretches, when it is, because of some overriding purpose in the writer's mind. Or, is it coherent because the writer is unconsciously constructing his text to obey certain long-range statistical rules? And, we wonder, how many times does a person let his pen or tongue, started by some initial impetus, merely run through a sequence of probable words?

This sort of investigation became interesting for its own sake. A couple of hours spent in a conference room with two mathematicians and two engineers produced a half a dozen curious forty-word bits. It is scarcely worthwhile to quote the whole of these, but some selected sentences may be of interest.

* I quote from Menninger's "The Human Mind," third edition, page 233, "Have just been to supper. Did not knowing what the woodchuck sent me here. How when the blue blue blue on the said anyone can do it—?

"When cooked asparagus has a delicious flavor suggesting apples."

"No man should judge his actions by his wife Susie."

"It happened one frosty look of trees waving gracefully against the wall."

We see that the statistics involved are sufficient to give "meaning" frequently, but are scarcely adequate to insure "truth". But, if we mean by truth merely that which we are likely to find written in encyclopedias, statistics could presumably supply it, too. With the statistics which we have included, however, any merit of such compositions is more apt to be aesthetic than factual. The last sentence has, for instance, a rather pleasing twisting effect which might have escaped a conscious artist.

We are reminded that philosophers have argued for years about how much of art lies in the work of the artist and how much lies in the observer. I do not know whether or not Dr. Shannon had anything of this in mind, but these consequences of his work certainly have an interesting bearing on the matter. Here there is no creator or "artist." The structure of the words is based merely on statistics, or, on the likelihood of their occurring in a certain order. Yet, they may have "meaning" for the reader, and he may have an aesthetic appreciation of them.

The passages quoted above were rather disconnected. Our interest finally led us to try a drastic and unscientific experiment. If lack of long-range connection were the chief trouble with the text, we could remedy that. On the bottoms of the slips of paper on which we wrote the words, in plain sight of all, various subjects were indicated, among them *salaries, murder story* and *women.*

The statement on salaries is of interest for a certain partisanship:

"Money isn't everything. However, we need considerably more incentive to produce efficiently. On the other hand too little and too late to suggest a raise without reason for remuneration obviously less than they need although they really are extremely meager."

The *murder story* slip contained a passage which goes a little beyond the bloodiest and most disconnected of the genre:

"When I killed her I stabbed Paul between his powerful jaws clamped tightly together. Screaming loudly despite fatal consequences in the struggle for life began ebbing as he coughed hollowly spitting blood from his ears."

It was on the final slip, *women,* that chance really spoke through clearly. The forty-two word statement is succinct but not entirely quotable. The last sentence says a great deal:

"Some men repeat past mistakes again and again and again."

Perhaps this adage appeared because it has so likely a connection with any part of our lives, our scientific interests included.

THE END