[*Mechanical Translation*, vol.5, no.2, November 1958; pp. 67-73]

The Use of Statistics in Language Research

A. F. Parker-Rhodes, Cambridge Language Research Unit, Cambridge, England

The literature concerning the application of statistics to linguistic problems and in particular to mechanical translation is reviewed. The conclusion is that much of the work done is of little direct use for mechanical translation, and that some of it is based on a misapprehension of what statistical techniques can in fact do. Statistical methods can play a useful part in the development of mechanical translation procedures once these have been well established, but have little to contribute at the present stage of the work.

THERE ARE many ways in which statistical techniques might be pressed into the service of language research, and in particular the theory of mechanical translation and information retrieval. Most of these have had their advocates, The purpose of this paper is to review briefly the literature of the subject, and to draw conclusions as to how much of this work can be regarded as a legitimate use of statistics, and as to how relevant it is to the progress of language-processing technology.

There appear to be five main topics covered. First, I shall enumerate these, and then I shall refer seriatim to the works available in the C.L.R.U. library upon each of them. 1) Lexicography: this includes the methods and techniques of compiling lexical information, whether this takes the form of a dictionary of a more or less conventional character, or a thesaurus. 2) Approximative Methods: these are methods of machine translation which aim to rely on keeping errors below a preconceived threshold of tolerance; they use statistics mainly to predict how little work need be done to achieve this. 3) Economics: included here are applications of statistics to ascertain the size of computers needed, the time taken to operate programs, etc. 4) Coding: the problems of coding of information have a statistical aspect whenever code-compression is employed. 5) Cryptography: a peripheral subject, but perhaps worth inclusion.

Applications to Lexicography

A good deal of theoretical work has been done on statistical techniques of a kind which could or might be applied to the study of word frequency. The general problems are of a kind of frequent occurrence in biology, and so have received some attention from that quarter. Of this general kind is the work of Good.[1] More specifically concerned with language problems are the contributions of Mandelbrot [2,3] on word-frequencies. This author points out that a knowledge of word-frequency distributions could be useful to the lexicographer, but he is not himself concerned to make this application. In fact, no one seems to have done so, except Koutsoudas,[4] who in fact concludes that the so-called Zipf and Joos laws are insufficient to give reliable predictions of the size of dictionaries needed in machine translation, and consequently recommends the accumulation of further empirical material with this end specifically in view.

[1] I. J. Good and G.H.Toulmin, "The number of new species and the population coverage, when a sample is increased, " Biometrika, 43, pp. 45-63 (1956).

[2] B. Mandelbrot, "Linguistique statistique macroscopique: Theorie mathematique de la loi de Zipf," Institut Henri Poincare, Seminaire de Calcul des Probabilites, (June 13, 1957).

[3] B. Mandelbrot, "Structure formelle des textes et communication," Word, 10, pp. 1-27 (1954).

[4] A. M.Koutsoudas and R.E. Machol, "Frequency of occurrence of words; a study of Zipf's law with application to mechanical translation, " University of Michigan, Engineering Research Institute, Publication 2144-147-T (1957).

Koutsoudas' statistical techniques are apparently adequate for his purpose, and he has compiled the required data and analyzed them. No one else has apparently taken statistical methods as seriously as this, and most references to the subject merely suggest that an application of statistics to dictionary making should be made,[5] or even in one case that no dictionary could be made without previous statistical analysis.[6]

The use which most of these authors have in mind is to find out how large a dictionary must be in order to contain, with a given fiducial probability, all the words of particular kinds of text. A secondary application is in finding some way of arranging the entries of a dictionary which will reduce searching time by making the most frequent words come up before the less frequent ones. Much more sophisticated is the idea behind compiling a thesaurus. In a thesaurus we have not merely a list of words with coded information upon them, but a mathematical system whose elements represent sets of words, so arranged that, ideally, every word in the system can be defined by listing the sets in which it occurs. If this were done properly, it should be possible to find a word, or at least most words, by specifying not all the sets in which it occurs, but only some of them; thus, it might be possible to specify a set of sets by considering the context of a given word, as well as itself, which would be enough to identify the given word as exactly as we might wish, provided our thesaurus contained enough information suitably organized.

Obviously, the success of such a scheme is a matter which could be statistically assessed, and in some measure no doubt statistically predicted. Thus, those who have considered the use of a thesaurus in MT have not been slow to appeal to statisticians for help in the very considerable labor of compilation involved. However, in fact, they have not progressed very far. As Luhn[7] puts it, "the formation of notational families (his name for thesaurus heads) is a major intellectual effort, to be undertaken by experts familiar with ....the special field of the subject-literature." This major effort has to be done before one can begin to apply one's statistical methods; Luhn himself makes no pretence of actually doing any statistics. On the other hand Gould, [8] who also considers thesaurus methods, presents the appearance of statistical computation. His problem is the translation of Russian mathematical texts into English, and he is concerned to assess the magnitude of the problem of 'multiple meaning' by statistical means. He defines an 'index of multiplicity' in algebraic formulae, and evaluates it for various word-classes (according to the system of Fries [9]), and presents numerical tables of the result. Actually the figures are not statistical in the strict sense, since no significance tests are done (nor is it shown that his index is a sufficient statistic), and the tables only show such facts as, for example, that prepositions are particularly liable to have multiple meanings. It cannot therefore be said that Gould's use of figures has added to what a discursive argument could have more lucidly put across.

One must conclude, from the few attempts which have been made actually to use statistics for lexicographic purposes, that in this field, a valid application exists only after the lexicographic data have been compiled. The same is true, whether the compilation takes the form of a dictionary or a thesaurus. Given these data, one can assess its adequacy, and even propose specific improvements of a major or minor kind, as a result of statistical analysis of its performance. But before the lexicographer has done his work, the statistician has nothing to use as data.

Approximative Methods

One answer to the difficulties raised by the attempt to reduce translation to a mathematically definite procedure is to base one's procedure on the opposite conception, namely that

[5] N. Chomsky, Syntactic Structures, Mouton and Company, The Hague (1957).

[6] V.A.Oswald and S.L.Fletcher, "Proposals for the mechanical resolution of German syntax patterns," Modern Language Forum, vol. 36, no. 3-4.

[7] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information, " IBM Journal of Research and Development, vol.1, no.4, pp. 309-317 (Oct. 1957).

[8] R.Gould, "Multiple correspondence," MT, vol. 4, no. 1/2, pp. 14-27 (Nov. 1957).

[9] C. C. Fries, The Structure of English, Harcourt, Brace and Company, New York (1952).

that instead of mathematical definiteness one should aim at acceptable approximation to the best that a human translator can do. In that case, it becomes important to know how much work must be directed to removing the errors present in too crude a procedure, in order to reduce the remaining errors to a point below some given threshold of tolerance. This is a statistical problem familiar in industry and in military applications. There seems good reason to expect that, if the approximative approach to MT is accepted as a useful one, it will rest largely on a statistical foundation.

A good example of the kind of work which is relevant to this viewpoint is that of Yngve[10] on 'gap analysis'; even though this is not oriented directly to MT application. This aims to supplement syntactic analysis of a text by a statistical procedure designed to reveal discontinuities between pattern-groups (of words) previously established by analysis of a sufficiently large corpus of texts. Insofar as the results of such analysis can be regarded as an acceptable model of actual linguistic analysis, the procedure is perfectly sound and, it must be admitted, highly ingenious. It is not like the deceptive figuring which we sometimes meet under the guise of statistics in language research. Most often, however, approximative methods are directed to eliminating errors of a lexicographic kind. For example, Glazer[11] has tried to work out the statistics necessary to permit the insertion of English articles into a translation from the Russian. He makes no great claims for the result but it is at least apparent from his work that the amount and detail of the statistical information required to 'solve' this problem, even within the framework of the approximationist philosophy, would be very considerable. In fact, it is unclear why it should be supposed any 'easier' than using real linguistics to do the job.

A better case is made out by King and Wieselman,[12] who have made some useful estimates of the work involved in progressively improving a crude translation by replacing more probable (and thus sooner tried) renderings of a given word or phrase by successively less probable ones. Once again, the conclusion seems to be that an acceptable amount of computation work leads to a still unacceptably erroneous result, though this no doubt depends on the purpose governing our choice of method.

The nature of approximative methods of trans lation is seen at its clearest when the attempt is made to get at the true meaning of a word by comparing it with successively wider areas of 'context.' The idea is that if the word itself is not sufficiently determinate to be translated by one-one equivalence, it may be that comparing it with the next word, or the last word, will suffice to reduce its possible equivalents to one failing that, we try two neighboring words, and so on till the desired result is achieved. This of course is a very crude model of what context really is, and, as I have stated it, depends on the untenable view that each word has a definite number of 'meanings', one of which has to be selected as its translation in the given context. These are just the assumptions made by Kaplan, [13] who made a statistical study of the problem; he collected his data by asking human informants to write down how many 'meanings' of selected words occurred to them, when the said words were presented in company with var ying numbers of neighboring words. His conclusions were not very detailed, largely becaus his informants were too few to provide a really adequate sample, but they showed clearly enough that indeterminacy of meaning was a decreasing function of size of context. There would be scope for a similar study, on a larger scale and with more powerful statistical methods, using a realistic model of what constitutes context and a realistic measure of the indeterminacy of semantic content; this would however be difficult to do. Like most applications of statistics to MT it would only really give useful results when applied to an already mechanized translation procedure. It would be far too slow and laborious to constitute an aid to constructing a mechanized procedure.

[10] V. H. Yngve, "Gap analysis and syntax," Transactions IRE, vol.IT-2, no. 3, pp. 106-112.

[11] S. Glazer, "Article requirements of plural nouns in Russian chemistry texts," Georgetown University, Institute of Languages and Linguistics, Seminar Work Paper MT. 42 (1957).

[12] G. W. King and I. L. Wieselmann, "Stochastic methods of mechanical translation," MT, vol. 3, no. 2, pp. 38-39 (Nov. 1956).

[13] A. Kaplan, "An experimental study of ambiguity and context," MT, vol. 2, no. *2,* pp. 39-46 (Nov. 1955).

Application to the Economics of Language Processing

It may be objected that it is still much too early to embark on a serious study of the economic aspects of MT. It is necessary, however, from time to time to reassure those concerned that the scale of the enterprise is not wholly disproportionate to the sums which its ultimate users will be prepared to devote to the necessary equipment. It can hardly be said that adequate data yet exist on which to base an informed answer to the question, "How big a computer must one have to do mechanical translation properly?" The question is of course a statistical one and in this sense is relevant to the present enquiry but it need not detain us long. Several workers have referred to the problem, but only Yngve[14] has given any detailed estimates. Their worth is somewhat dependent on accepting a particular view of the nature of the MT procedure but may be accepted to an order of magnitude, at least until more substantial data are available.

Coding and Code Compression

In large measure the coding problems arising in MT and in library work are the same as those occurring in other branches of communication engineering. The need for code compression perhaps arises more urgently in MT, because of the great bulk of the material to be stored, but the mathematical problems it presents are the same as in other fields, except where, as in the use of thesaurus methods, the mathematical structure of the information to be coded imposes special restrictions.

I do not intend to refer to the already considerable literature on code compression. Specific applications to MT have been discussed by Mooers.[15] This work however depends on using a tree-type semantic classification, as has hitherto been done in most information retrieval systems. The statistics of the process would be appreciably different in a lattice system.

Less specific to our immediate subject are the methods, many of them well known, for compressing alphabetic codes. Quite powerful methods are possible here because of the very great redundancy in alphabetic writing. They are discussed, in general terms and without statistical analysis, by Mukhin[16] and Panov.[17] In general it may be said that none of this work is either controversial or novel; but the statistics of code compression in thesaurus systems is still (as far as published work goes) an unexplored field.

Cryptography

As for coding problems, there is a large literature on cryptography and code design which I do not intend to explore. There are however some special points of contact between cryptography and language research in which statistics could play a part. Yngve[18] has written an interesting paper in which he treats of the translation problem (especially translation out of unknown languages) as a special case of the problem of decoding a message without the advantage of a complete code-book to do so. The approach potentially involves the use of statistics, and, while Yngve does not carry the analysis far enough to make actual calculations it is clear that this could be done. The difficulty is that the analogy between translation and the decipherment of a coded message is really more metaphorical than strictly formal. It is therefore unclear how far the results of such investigations will really be relevant.

General Commentary

Of the two main ways in which statistics can be applied to scientific enquiry, the observational and the predictive, only the first has

[14] V. H. Yngve, "The technical feasibility of translating languages by machine," Transactions AIEE, Paper 56-928 (1956).

[15] C. N. Mooers, "Zatocoding and developments in information retrieval," Aslib Proceedings, vol. 8, pp. 3-19 (1956).

[16] I. S. Mukhin, An Experiment in Machine Translation Carried out on the BESM, Academy of Sciences of the USSR, Moscow (1956).

[17] D. Panov, Concerning the Problem of Machine Translation of Languages. Academy of Sciences of the USSR, Moscow (1956).

[18] V. H. Yngve, "The translation of languages by machine," Information Theory, (Third London Symposium), Butterworth's Scientific Publications (London), pp. 195-205.

really been explored in our field. Observational statistics requires that there be a population of entities of which we cannot hope to acquire a complete knowledge, although we can obtain such knowledge of small samples of the population. These samples have to be taken subject to certain rather rigid precautions and in most statistical work are either created by carefully designed experiments or obtained by properly planned observations on the population as it exists in nature.

In the lexicographic applications these prerequisites are not very well met. When the population is the words in a dictionary, it is not a population of which our knowledge is fragmentary in the sense required. On the contrary, we already know (or someone must know) everything about them that we shall ever discover by our analysis, else the dictionary could not have been written. When the population is composed of words in a text, we are in no better position, for although here a real population exists, we either sample the whole population, in which case what we do is not really statistics but census-taking, or we postulate the existence of a population of which our text is a sample. This is in fact what most of the workers along this line appear to do, but it embodies a statistical fallacy, namely, that of creating a sample by definition. It is legitimate to define a population, ostensively or otherwise, and then set about obtaining samples from it, for then the legitimacy of the sampling procedure is open to test and discussion; it is not legitimate to ostend a sample and say "let there be a population of which this is a sample," for then there is no sampling procedure, and the assumptions of probability theory, on which the analysis of the results must be based, will not be correct.

The same objection does not apply to the application of statistics to the study of approximative methods of translation. Here the criticism which suggests itself, against all the work in this field, is the very artificial character of the systems studied. One feels it would hardly be worth while to do very much calculation on such systems. In fact, hardly any has been done. Many have said that they recognize the problem as statistical, but even those who, like Kaplan,[13] actually set out figures do not actually subject them to real statistical analysis. The application of statistics to these approximative methods is still more a potentiality than a fact.

This indeed is largely true of the whole field. There has been far more written about statistical work in translation and information retrieval than actual work done. Apparently no one has yet clearly stated the very limited nature of the applications possible, but many have borne witness to it by inaction. Broadly speaking, the populations which it would be valuable to have information upon are those provided by mechanically translated texts themselves, and the reason that we want to have the information is so as to be able to spot what is wrong with the translation procedure used. Human texts are not suitable material for the statistician because the information we can hope to get from them is either already available or is more efficiently extracted by the methods of the linguist than by those of the statistician.

The indeterminacy which does exist in language is the indeterminacy which arises from the mapping of a continuous territory onto a chart with a finite resolving power; it is not the result of an intrinsically indeterminate use of a discrete set of symbols however complicated. This being so, language can certainly be described in statistical terms. But there is no point in describing it, because the object of the translator (human or mechanical) is instead to use it, in the same sense that one uses a mathematical system to calculate with. Since we shall never do this 'perfectly,' it will always be worth while to estimate the gravity of our failures and this will be a large enough field for the statistician for a long time. But this activity will only begin when the output of failures becomes copious enough to provide the statistician with large populations and the opportunity of applying proper sampling methods to them. This has not yet happened.

Many of those who have written on this subject seem to have the unexpressed belief that there is in language, or our use of it, something essentially indefinite which can be dealt with mathematically only in statistical terms. If this were so, the conveyance of precise information by talking would be impossible. To some extent the area of possible meanings of a remark can be regarded as a probability distribution, but it is of the kind that is almost everywhere zero and has a finite value only within a restricted region. If we deal in 'areas of meaning' instead of in point-like 'right' and 'wrong' meanings, there are indeed definite rules which tell us what remarks do <u>not</u> mean. Deliberately ambiguous statements can be made in all languages, but even these can be recognized as such by the rules. The problem for the translator is to find out the rules of the languages concerned and to apply them. It is conceivable that this is too difficult for a machine to do; in that case, perhaps a statistical approximation to the desired translation would be a next-best. But it is a substitute, not the real thing.

This paper was written with the support of the National Science Foundation, Washington, D. C.

The following comments were received from people whose work is mentioned in the preceding article. These comments are published with the permission of those concerned.

I agree with the point of view expressed in this paper by Parker-Rhodes, but I fail to see the relevance that he notes of my work on gap analysis to the approximative approach to MT. The gap analysis procedures were intended as a tool for the linguist who wants to discover non-approximative methods in MT.

I would like to see a clear distinction made between analysis of a language for the purpose of deducing its rules or structure, and analysis of a sentence to obtain its structure for possible use when translating it by machine. We may not be able to mechanize the former as easily as the latter. These two kinds of analysis are as different as the science of chemistry, aiming to discover the general laws of chemical composition and reaction, and the analysis of an unknown compound of mixture for its ingredients and their mode of combination.

V. H. Yngve

Footnote 5, and the accompanying sentence in the text (page 2, second paragraph) should be deleted, as factually inaccurate. No such statement is made in Syntactic Structures. Statistics is discussed only on pp. 16,17, — lexicography is not mentioned at all.

Noam Chomsky

I am sorry to say that the wide range of items covered by Parker-Rhodes and the (to me) excessive economy of words made it difficult to follow him in several places, including the section where he deals with my own piece on "Article Requirements of Plural Nouns in Russian Chemistry Texts."

Frankly, I'm not sure that I understand what he is objecting to.

He did not challenge the accuracy or usefulness of the principle of article insertion I proposed or even fault the statistical methodology, as far as I could make out. May I add, for what it may be worth, that I submitted my paper in advance of delivery to a professor of statistics from Stanford, who found my approach wholly acceptable. In the semi-public demonstration of the Lukjanow code-matching technique held in Washington on August 20th, the percentage of correct article placement (in some 300 sentences, including those in the random text) tallied perfectly with the percentage mentioned in my paper. Parker-Rhode's statement "It is unclear why it should be supposed any 'easier' than using real linguistics to do the job" (p. 6) is particularly baffling. Since the article study originated with and was based wholly on an analysis primarily of English usage and possible Russian morphologico-syntactic decision points, and various counts made afterwards only to ascertain whether the formulation provided "useful" predictability, the implication that the tail wagged the dog is certainly unwarranted.

It was not my intention to use statistics to "solve" the problem; rather to indicate that the formulations suggested permit mechanical insertion or omission of articles with a fairly high degree of accuracy. I can't see how statistics as such are useful in MT except as indicators of the validity of a proposed solution.

In my view there is no single solution of a foreign text. Some 15 years experience as a translation editor, translator (both of scientific and purely literary works), and student of the art of translation have led me to believe that there are likely to be as many versions or solutions of a text (with varying quality, of course) as there are translators. The acceptability of a given translation rests with the individual reader whose reactions are dictated by his background knowledge of the Subject, sensitivity to the nuances of his native language, and the use to which he intends to put the translation. That is why I am a proponent of "approximationism" in language which I think reflects the reality of the human potential, however weak, rather than the ideal, however desirable.

What is needed now as far as the articles are concerned is not more statistical information per se but greater insight into the way they are behaving today. As you know, English article usage has been evolving over a long period of time and the process is far from complete. Under the present influence of the radio and, particularly, the press, with its emphasis on conciseness, there seems to be a trend away from the article in certain types of constructions, e.g. with abstract nouns in possessive phrases. Elsewhere speakers not infrequently have a choice between "a" and "the", etc., with faint semantic or even idiomatic difference between either. How much precision can we (or should we try to) build into a /the translation machine ?

Sidney Glazer

Dr. Gould's untimely and tragic death in the Alps last summer precludes a personal comment on his part. I feel sure, however, that he would wish simply to let his published work speak for itself.

Anthony G. Oettinger