REPORT ON THE STATE OF MACHINE TRANSLATION IN THE UNITED STATES AND GREAT BRITAIN

Yehoshua Bar Hillel

Hebrew University

TECHNICAL REPORT NO. 1

THIS REPORT WAS PREPARED FOR THE UNITED STATES OFFICE OF NAVAL RESEARCH INFORMATION SYSTEMS BRANCH, UNDER CONTRACT NO. NONR-2578(00). N.R. 049-130

Jerusalem, Israel
15 February 1959

SUMMARY

The present report contains a critical survey of the status of research on the mechanization of translation in the United States and Great Britain, especially of the methodological and linguistic aspects of this problem. The unreasonableness of aiming at fully-automatic high-quality translation is stressed, the shortcomings of the approaches sticking to this aim discussed, and its lowering to that of either providing machine aids to high quality translation or providing fully-automatic but low-quality translation advocated. Some proposals for the organization of further research and for the improvement of cooperation are made.

The four appendices contain some statistics and three talks, in various stages of publication, before international meetings on matters of principle of direct or indirect concern to MT.

CONTENTS

1. Machine translation (MT) has become a multi-million dollar affair. It has been estimated[1] that in the United States alone, something like a million and a half dollars were spent in 1958 upon research more or less closely connected with MT with approximately one hundred people, among them sixty with M.A., M.Sc. or higher degrees, working in the field. No comparable figures are available for Russia[2], but it is generally assumed that the number of people engaged there in research on MT is higher than in the States. There exist, in addition, two centers of research in MT in England, a third being in the process of formation. Outside these three countries, MT has been taken up only occasionally, and no additional permanent research groups seem to have been created. Altogether, I would estimate that the equivalent of between 200 and 250 people were working full-time on MT at the end of 1958, and that the equivalent of three million dollars were spent during this year on MT research. In comparison, let us notice that in June 1952, when the First Conference on Machine Translation convened at M.I.T., there was probably only one person in the world engaged more than half-time in work on MT, namely myself. Reduced to full-time workers, the number of people doing research on MT could not at that time have been much more than three, and the amount of money spent that year not much more than ten thousand dollars.

For this conference, I had prepared in mimeograph a report on the state of the art which was later published in print in American Documentation[3]. That report was based upon a personal visit to the two or three places where research on MT was being conducted at the time, and seems to have been quite successful, so I was told, in presenting a clear picture of the state of MT research as well as an outline of the major problems and possibilities. It has been my feeling that the time has arrived to critically evaluate the progress made during the seven years that have since passed in order to arrive at a better view of these problems and possibilities. To my knowledge, no evaluation of this kind exists, at least not in English. True enough, there did appear during the last year two reviews of the state of MT, one prepared by the group working at RAND Corporation[4], the other by Martin H. Weik and George W. Reitwiesner at the Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland[5]. The first of these reviews was indeed well prepared and is excellent as far as it goes. However, it is too short to go into a detailed discussion of all existing problems, and, in addition, is not always critical to a sufficient degree. The second review seems to have been prepared in a hurry, relies far too heavily on information given by the research workers themselves, who by the nature of things will often be favorably biased towards their own approaches and tend to overestimate their own actual achievements, and does not even attempt to be critical. As a result, the picture presented in this review is somewhat unbalanced, though it is still quite useful as a synopsis of certain factual bits of information. Some such factual information, based exclusively upon written communication from the research groups involved, is also contained in a recent booklet published by the National Science Foundation[6]. Brief histories of MT research are also presented in the Introductory Comments by Professor Léon E. Dostert to the Report of the Eighth Annual Round Table Conference on Linguistics and Language Study[7] as well as in the Historical Introduction to the recent book by Dr. Andrew D. Booth and associates[8].

The present report is based upon personal visits to almost all major research centers on MT in the United States, the only serious exception being the center at the University of Washington, Seattle, and upon talks with members of the two research groups in England, as well as, of course, upon a study of their major publications including also, as much as possible, progress reports and memoranda. In addition, a circular letter was sent to all research groups in the United States asking for as detailed information as possible concerning the number and names of people engaged in research within these groups, their background and qualifications, the budget, and a short statement of the plans for the near future. I did this in order not to be forced to rely too heavily on my memory, and also in view of the fact that I usually did not ask during my visits for these particulars, in order to save the short time that stood at my disposal for discussion of things like the approach adopted and the methods used, a clear view of which could not be obtained by simple clerical operations. However, not all groups were equally responsive to this circular letter, and the factual information of those groups which did not supply the requested information is therefore based upon memory, aided, of course, by information obtained from a perusual of the mentioned reports and reviews as well as from other sources. No very high degree of accuracy is aimed at in respect to this kind of information, the major aim of the present report being rather a discussion of approaches, methods, problems and possible solutions. It was felt, however, that an up-to-date and reasonably accurate picture of the quantity and quality of the manpower engaged at present in the research on MT, as well as an estimate of their financial resources would add an aspect of some value to the overall picture. This kind of information will therefore be presented in Appendix I in tabular form.

The visits upon which this report is based were made in October and the first week of November, 1958. The circular letter was sent on the 20th of November, 1958, and the answers, if any, received during December 1958 and January 1959. On November 13, I reported before a group of representatives of various government and military agencies upon the impressions obtained during my visits, and promised to supplement this oral report by a written one as quickly as possible. The present document is the fulfillment of this promise.

2. During the first years of the research in MT, a considerable amount of progress was made which sufficed to convince many people, who originally were highly skeptical, that MT was not just a wild idea. It did more than that. It created among many of the workers actively engaged in this field the strong feeling that a working system is just around the corner. Though it is understandable

that such an illusion should have been formed at the time, it was an illusion. It was created, among other causes, also by the fact that a large number of problems were rather readily solved, and that the output of machine-simulated "translations" of various texts from Russian, German or French into English were often of a form which an intelligent and expert reader could make good sense and use of. It was not sufficiently realized that the gap between such an output, for which only with difficulty the term "translation" could be used at all, and high-quality translation proper, i.e., a translation of the quality produced by an experienced human translator, was still enormous, and that the problems solved until then were indeed many but just the simplest ones, whereas the "few" remaining problems were the harder ones -- very hard indeed.

I am not sure whether there still exist many groups which think that fully-automatic, high-quality machine translation (FAHQMT)[9)] is attainable in the near future, say within five years or so. Claims to this effect have been made by one of the four subgroups working on MT at Georgetown University. I shall discuss these claims below. But let me state already at this point that I could not be persuaded of the validity of these claims. On the contrary, I am quite ready to commit myself to concoct Russian sentences or, should this for some reason be regarded as unfair, to exhibit actually printed Russian sentences for which a perusual of the program of this group, or of any other group that would offer in the near future a method of fully-automatic MT, would result either in gibberish or, what is even worse, in meaningful but wrong translations. Appendix IV, a paper submitted to the International Conference on Information Processing, Paris, June 1959, contains a proof of the non-feasibility of FAHQMT.

Most groups, then, seem to have realized that FAHQMT will not be attained in the near future. Two consequences can be drawn from this realization. One can stick to the aim of FAHQMT and be ready to renounce attainability in the near future, perhaps even attainability altogether, but still hope that in the pursuit of this aim interesting theoretical insights will be gained which will justify this endeavor, whether or not these insights will ever be exploited for some practical purpose. Or one may insist on sticking to attainability in the near future but then be ready to give up the ideal of FAHQMT, and be satisfied with some less ambitious scheme. Both conclusions are equally valid but should lead to rather different approaches. Unclarity in this respect, vague hopes that somehow or other both aims can be attained simultaneously and by the use of the same methods, must lead to confusion and result in waste of effort, time and money. If one is interested in MT as a practical device, meant to reduce the existing heavy load of valuable texts in foreign languages waiting to be translated into English, he must realize that some compromise is absolutely necessary. There are two possible directions in which such a compromise could go: one could sacrifice quality or one could reduce the self-sufficiency of the machine output. There are very many situations where less than high-quality translation is satisfactory. There is no need to present examples. If, however, high-quality is mandatory -- and I do not think, for instance, that scientists are prepared to be satisfied with less than the present average standard of human translation, while many regard this standard as too low for their purposes --, then the machine output will have to be post-edited, thereby turning, strictly speaking, machine translation into <u>machine aids to translation</u>.

3. In the remainder of this report, I shall exclusively deal with those situations where translation involved has to be high-quality. It should be easy to see how the conslusions at which I arrive have to be modified in order to deal with situations in which lesser quality is satisfactory.

As soon as the aim of MT is lowered to that of a <u>machine--post-editor partnership</u>, the decisive problem becomes to determine the region of optimality in the continuum of possible divisions of labor. It is clear that the exact position of this region will be a function of, among other things, the state of linguistic analysis to which the languages involved have been submitted. It may be safely assumed that, with machine-time/efficiency becoming cheaper and human time becoming more expensive, continuous efforts will be made to push this region in the direction of reducing the human element. However, there is no good reason to assume that this region can be pushed to the end of the line in the near future.

It seems that with the state of linguistic analysis achieved today, and with the kind of electronic computers already in existence or under construction, especially with the kind of large-capacity, low-cost and low-access-time internal memory devices that will be available within a few years, a point has been reached where commercial man-machine translation outfits could become a practical reality. This, however, is still conditioned by the two following factors: first, a reliable and versatile mechanical print reader will have to be available. It has been estimated that the cost of retyping printed Russian material into a form and on a medium that could be processed by a machine would amount, under present conditions, to about one fourth of a cent per word[10]. This estimate is probably much too low, as the quality of the retyping has to be exceptionally high, in order to avoid printing mistakes which would probably be quite harmless for a human reader but could be rather disastrous for machines which so far are totally unable to deal with misprints. The original text would therefore have to be keypunched by two operators, verified, etc.[11] The difference between one half of a cent per keypunched word and, say, one twentieth of a cent per print-read word[12] could make all the difference. Secondly, a concerted effort will have to be made by a pretty large group in order to prepare the necessary dictionary or dictionaries in the most suitable form. In addition, a good amount of thinking accompanied by an equally large amount of experimenting will still have to go into the determination of the location of the interval in the above mentioned continuum within which the optimal point of the division of labor between machine and post-editor will have a good chance of being situated, as a function of the specific translation program and the specific qualities of the envisaged post-editor. Among other things, these studies would have to determine whether some minimal pre-editing, while requiring but very little knowledge of the source language by the pre-editor could not be utilized in order to reduce the load of the machine by a considerable amount[13]. This is just one of the very many points which have still to be settled before MT is in business. Another point which has not been treated so far with sufficient incisiveness, mostly because the ideal of fully-automatic translation diverted the interests of the research workers into other, less practical directions, is the old question whether MT dictionaries should contain as their source language entries all letter sequences that may occur between spaces, sometimes called alternate words, or rather so-called base words[14], or perhaps something in between. This question is clearly highly dependent, among other things, upon the exact type of internal and external memory devices available, and it is therefore mandatory to have a reliable estimate of this dependence. It is obvious that the speed of the machine part of the translation, and thereby the cost of the total translation process, will depend to a high degree on the organization of the dictionaries used. Most workers in the field of MT seem to have rather definite, though divergent, opinions in this respect. However, I am not aware of any serious comparative studies, though the outcome of such studies most surely will have a considerable impact upon the economics of MT.

In general, the intention of reducing the post-editor's part has absorbed so much of the time and energy of most workers in MT, that the problem of whether partially-automatic translation, even with such a large amount of participation on behalf of the post-editor as would be required under present conditions, is not nevertheless a desirable and feasible achievement has not received sufficient discussion. I fully understand the feeling that such an achievement is not of very high intellectual caliber, that the real challenge has thereby not yet been taken up, but I do not think that those agencies, for whom any reduction of the load imposed at the moment on the time of highly qualified expert translators is an important achievement, should necessarily wait with the installation of commercial man-machine translation outfits until such a time when the post-editor's part has become very small, whatever the amount of satisfaction the MT research worker will get from such an achievement.

4. At this stage, it is probably proper to warn against a certain tendency which has been quite conspicuous in the approach of many MT groups. These groups, realizing that FAHQMT is not really attainable in the near future so that a compromise is definitely indicated, had a tendency to compromise in the wrong direction for reasons which, though understandable, must nevertheless be combatted and rejected. Their reasoning was something like the following: since we cannot have 100% automatic high-quality translation, let us be satisfied with a machine output which is complete and unique, i.e., a smooth text of the kind you will get from the output of a human translator (though perhaps not quite as polished and idiomatic), but which has a less than 100% chance -- I shall use the expression "95%" for this purpose, which is, of course, not to be taken literally -- of being correct. Such an approach would be implemented by one of the two following procedures: the one procedure would require to print the most frequent target-language counterpart of a given source-language word whose ambiguity has not been resolved by the application of the syntactical and semantical routines, necessitating, among other things, large-scale statistical studies of the frequency of usage of the various target-renderings of many, if not most, source-language words; the other would work with syntactical and semantical rules of analysis with a degree of validity of 95%, if only this degree were sufficient to insure uniqueness and smoothness of the translation.

I regard this approach as wrong and, more than that, dangerous, so long as high quality is essential. Since so many sentences, "5%" of a given text, will have a good chance of being mistranslated by the machine, it is by no means clear whether the post-editor will be able to correct these mistranslations, especially in view of the fact that the machine output is so smooth and grammatical (so let us assume for the sake of the argument, though I doubt it whether even this much can really be achieved at this stage of the game) that he might be able to find only few cues to warn him that something is wrong with it. It is not inconceivable that the machine translation would be so wrong at times as to lead its user to actions which he would not have taken when presented by a correct translation. (When I am talking about "100%", I obviously have in mind not some heavenly ideal of perfection, but the product of an average qualified translator. I am aware that such a translator will on occasion make mistakes and that even machines of a general low-quality output will not make some of these errors. I am naturally comparing averages only.)

But there is really no need at all to compromise in this direction of reducing the reliability of the machine output. True enough, a smooth machine translation looks impressive, especially if the reader is unable to realize at first sight that this translation is faulty ever so often, but this esthetically appealing feature should not blind us to see the dangers inherent in this approach. Since the post-editor will have to be involved at any rate in order to correct the machine's mistakes (and, I am quite sure, to do all kinds of other things, too), I regard it as much safer to compromise in the other direction. Let us be satisfied with a machine output which will ever so often be neither unique nor smooth, which ever so often will present the post-editor with a multiplicity of renderings among which he will have to take his choice, or with a text which, if it is unique, will not be grammatical. On the other hand, whenever the machine output is grammatical and unique it should be, to adopt a slogan used by Professor Anthony G. Oettinger, "fail-safe" (to about the same degree, to make this qualification for the last time, as the average qualified human translator output is fail-safe). Let the machine by all means provide the post-editor with all possible help, present him with as many possible renderings as he can digest without becoming confused by the embarras de richesse -- and here again we have quite a problem of finding an interval of optimality -- but never let the machine make decisions by itself on purely frequential reasons even if these frequencies can be relied upon. If these frequency counts could be done cheaply -- and I doubt very much whether this is feasible for such a high degree of reliability as would probably be required for our purposes -- , let this information too be given the post-editor. I am reasonable sure, however, that this additional information is not worth the enormous effort in time and money that would be required to obtain it under presently available methods, and that, in any case, MT should not wait until this information is obtained.

5. Let me finish this part of the report by warning in general against overestimating the impact of statistical information on the problem of MT and related problems. I believe that this overestimation is an outcome of the time, six or seven years ago, when many people thought that the statistical theory of communication would solve many, if not all, of the problems of communication. Though it is often possible by a proper organization of the research effort to get a certain amount of statistical information at no great extra cost, it is my impression that much valuable time of MT workers has been spent on trying to obtain statistical information of no possible impact on MT at all. It is not true that every statistic on linguistic matters is automatically of importance for MT so that the gathering of any such statistics could be regarded as an integral part of MT research without any need for additional justification.

Gathering of statistics is regarded by many MT groups as being part of a more general methodological approach -- the so-called "empirical approach"[15]. Adherents of this methodology are distrustful of existing grammar books and dictionaries, and regard it as necessary to establish from scratch the grammatical rules by which the source-language text will be machine analyzed, through a human analysis of a large enough corpus of source-language material, constantly improving upon the formulation of these rules by constantly enlarging this corpus. With regard to dictionaries, a similar approach is often implemented and a dictionary compiled from translations performed by bilingual members of the group or by other human translators considered to be qualified by this group. I regard this approach as unnecessarily wasteful in practice and as insufficiently justified in theory. It seems that the underlying distrust has been caused by the well-known fact that most existing grammars are of the normative type, hence often of no great help in the analysis of actual writing (and to an even higher degree, of actual speech), and that existent dictionaries are of such a nature that quite often none of the presented target-language counterparts of a source-language word are satisfactory within certain contexts, especially with regard to terms used in recently developed scientific fields. However, even in view of these facts, I believe that the baby has far too often been thrown away with the bath water. No justification has been given for the implicit belief of the "empiricists" that a grammar satisfactory for MT purposes will be compiled any quicker or more reliably by starting from scratch and deriving the rules of grammar from a large corpus than by starting from some authoritative grammar and changing it, if necessary, from observations of actual texts. The same holds <u>mutatis mutandis</u> with regard to the compilation of dictionaries. Not only has no justification been given or even seriously attempted, I think there are very good reasons to believe that no such justification can be given. Grammars have in general not wholly been dreamt up, nor have dictionaries been compiled by some random process. Existing grammars and dictionaries are already based, though admittedly not wholly, upon actual texts of incomparably larger extension than those that serve as a basis for the new compilers. Russian is not Kwakiutl, and with all due regard to the methods and techniques of structural linguistics and to the insights which this science has given us in respect to some deficiencies of traditional grammars, I do not think that it follows from their teachings that all existing codifications of languages with a highly developed literature should be totally disregarded. Let me add, without going here into details for lack of space, that the empiricalness of the derivations of grammar rules from actual texts is rather doubtful as such. For certain general methodological considerations one might as well be led to the conclusion that these rules incorporate a lot of subjective and highly biased and untested assumptions such that their degree of validity might very well, on the average, be lower than that of the well-established, often-tested and critically examined grammars, in spite of their normativity.

The only reasonable aim, then, for short-range research into MT seems to be that of finding some machine--post-editor partnership that would be commercially competitive with existing human translation, and then to try to improve the commercial effectiveness of this partnership by improving the programming in order to delegate to the machine more and more operations in the total translation process which it can perform more effectively than the human post-editor. These improvements will, of course, utilize not only developments in hardware and programming (especially automatic programming) of linguistical analysis, but also the experience gained by analyzing the machine output itself. Should it turn out that for the sake of competitiveness some use of a pre-editor, and perhaps even of a bilingual post-editor, would be at least temporarily required, then this fact should be accepted as such, in spite of the trivialization of the theoretical challenge of the MT problem which would be entailed by such a procedure.

6. It is now time to discuss in some detail the achievements of the various research groups in the United States and England.

Let me start with the two oldest groups, namely the group at the University of Washington, Seattle, headed by Professor Erwin Reifler -- to be referred to in the future as the Seattle group -- and the group at the Research Laboratory of Electronics at the Massachusetts Institute of Technology, Cambridge, headed by Professor Victor H. Yngve -- the MIT group. The reason for beginning with these groups is not so much their historical priority -- we recall that Reifler started his investigations into MT in 1949 and Yngve in 1953 when he took over from myself as I left MIT in order to return to the Hebrew University in Jerusalem -- but the fact that my personal contact with these groups during my visit in the States was either nil, in the case of the Seattle group, or very limited, in the case of the MIT group. In addition, the Seattle group seems to have published very little since the talk presented by Reifler before the Eighth International Congress of Linguists in Oslo, August 1957.[16] From a letter I recently received from Reifler, I understand that the Seattle group plans to publish very shortly a 600-page report summarizing in detail the total research effort of this group. In spite of considerable achievements in some highly specific problems such as the treatment of German compound words, which clearly pose a grave problem for MT with German as the source-language since this way of forming new German nouns is highly creative so that the machine would almost by necessity have to analyze such compounds before a dictionary look-up[17], it is not clear whether this group has been able to make great progress in the programming for complete syntactic resolution or in the solution of the problems posed by polysemy. Until October 1957, the Seattle group was concerned almost exclusively with determining the limits of attacking the MT problem by the use of lexicography alone. Only afterwards was it planned to deal with syntax and those aspects of semantics that cannot be solved by lexicography alone. I have no knowledge of the achievements made during the year and a half that have passed since. It is perhaps worthwhile to stress that this group does not adopt the "empirical approach" mentioned above, and is not going to be satisfied with so-called "representative samples", but is trying to keep in view the ascertainable totality of possible constructions of the source-language though representative samples are of course utilized during this process[18]. There is no need for me to stress at this point my agreement with this policy.

7. The MIT group, during the last years, has insisted on its adherence to the ideal of FAHQMT. For this purpose they regarded the complete syntactical and semantical analysis of both source and target-language to be a necessary prerequisite. It is, therefore, to these processes that their research effort has been mostly directed. It seems that this group is aware of the formidableness of its self-imposed task, and probably does not believe that even its prerequisite will be attained in the near future[19]. It believes, on the other hand, and I think rightfully, that the insights into the workings of language obtained by their research are valuable as such, and could at least partly be utilized in practical lower-aimed machine translation by whoever is interested in this latter aim. However, it will probably be admitted by this group that some of the research undertaken by it might not be of any direct use for practical MT at all. The group employs to a high degree the methods of structural linguistics, and is strongly influenced by the recent achievements of Professor Noam Chomsky[20] in this field.

Since the impact upon MT of Chomsky's recently attained insights into the structure of language is a controversial issue, it would have been worthwhile to spend here a few paragraphs on this point. However, since I presented my own views on this issue in a talk given before the Second International Congress on

Cybernetics, Namur, September 1958, scheduled to appear in the Proceedings of this Congress, as well as in a talk presented to the Colloque de Logique, Louvain, September 1958[21], which talks are reproduced here in Appendices II and III, I shall mention here only one point. The MIT group believes, I think rightly, that Chomsky has succeeded in showing that the immediate constituent model, which has so far served as the basic model with which structural linguists were working, in general as well as for MT purposes, and which, if adequate, would have allowed for a relatively simple completely mechanical procedure for determining the syntactical structure of any sentence in any language for which a complete description in terms of this model could be provided -- as I have shown already 6 years ago[22] -- is not fully adequate and has to be supplemented by a so-called transformational model. This insight of Chomsky explains also, among other things, why most prior efforts at the mechanization of syntactical analysis could not possibly have been entirely successful. The MIT group now seems to believe that this insight can also be given a positive twist and made to yield a more complex but still completely mechanical procedure for syntactical analysis. I myself am doubtful about this possibility, especially since the exact nature of the transformations required for an adequate description of the structure of English (or any other language) is at the moment still far from being satisfactorily determined. A great number of highly interesting but apparently also very difficult theoretical problems, connected with such highly sophisticated and rather recent theories as the theory of recursive functions, especially of primitive recursive functions, the theory of Post canonical systems, and the theory of automata (finite or Turing) are still waiting for their solution, and I doubt whether much can be said as to the exact impact of this new model on MT before at least some of these problems have been solved. I think that Chomsky himself cherished similar doubts, and as a matter of fact my present evaluation derives directly from talks I had with him during my visit.

The MIT group has, among other things, also developed a new program language which, though specially adapted for MT purposes, is probably also of some more general importance[23]. The fact that it was felt by this group that a program language is another more or less necessary prerequisite for MT is again the result of their uncompromising approach. To my knowledge no other group has been working in this direction, and the development of a program language is probably indeed not necessary, perhaps not even helpful for their restricted aims. I would, however, agree that a program language is indeed necessary for the high aims of the MIT group, though I personally am convinced that even this is not sufficient, and that this group, if it continues to adhere to its aims, will by necessity be led in the direction of studying learning machines. I do not believe that machines whose programs do not enable them to learn, in a sophisticated sense of this word, will ever be able to consistently produce high-quality translations.

8. In this connection, the work of the group at the University of Pennsylvania, Philadelphia, headed by Professor Zellig S. Harris should be mentioned. This group is wholly concerned with developing programs for the syntactical analysis of English, and is by no means directly concerned with the implications of its research for MT. They do, however, definitely hope that their research will lead to useful applications, not only for MT, but also for information retrieval and related problems. It is my painful duty to dispel at least some of these hopes. Though I think that the actual programs compiled by the Philadelphia group for the syntactic analysis of English embody solid achievements based upon valid intuitive insights as well as upon extremely painstaking and detailed observations, and in this respect equal if not superior to parallel achievements obtained during the same period by other groups concerned with the same problem (or rather, in most cases, with the materially different but

methodologically very similar problem of mechanically analyzing the structure of Russian, German, French, etc.), the theory behind these achievements seems to be of doubtful validity, if interpreted literally, and ill-formulated and misleading in any case. The detailed substantiation of this rather harsh judgment by quotations from the latest publication of this group, namely the paper[24] presented by Harris at the International Conference for Scientific Information, Washington, D.C., November 1958, will be undertaken elsewhere. One warning, however, is definitely indicated. The similarity of the terminology used by Harris and Chomsky is often deceptive. Chomsky, who is a former pupil of Harris and heavily indebted to him for many of the terms and underlying ideas, later came to use these terms in senses which are quite different from those given them by Harris. More strictly, whereas with Chomsky terms like 'transformation' or 'kernel' have pretty well determined senses, their vagueness not exceeding the usual range adhering to almost all scientific terms, they are not at all well-defined with Harris, and with him rely for their meaning on some far-fetched and under-developed analogy with the use of these terms in modern abstract algebra. In addition, Harris is much less cautious in his formulations than Chomsky. It is often quite certain that Harris could not possibly have intended to say what he seems to be saying if his words were taken literally. But even if the reader is armed with a high degree of good will, he is often at a loss how to interpret Harris' statements so as to save them from being patently false without becoming trivial. (I shall deal on still another occasion with the impact of Harris' ideas upon information retrieval. Let me say here only rather dogmatically that I regard this impact as being very slight, as far as one can judge at this time.)

I understand that the programming of the syntactical analysis of English is fairly advanced though it is apparently still very difficult to judge how advanced it is. In addition, this group is working with a Univac which does not seem to be a very effective machine for MT purposes.

In order not to be misunderstood, let me stress that my criticism refers only to Harris' description of what the process he calls "kernelization" is apt to achieve and that part of his theory of transformations which lies behind it. From a short discussion with him, I gathered that some of his formulations are indeed not to be understood literally, but I was unable to determine what exactly was left. It would be of some importance to get more clarity on this issue.

9. The largest group working on MT is that at Georgetown University, Washington, D.C., led by Professor Léon E. Dostert. The GU group comprises four subgroups. One of these is headed by Professor Paul L. Garvin and has been engaged during the last two years exclusively in the programming of the syntactical analysis of Russian. Their method seems to work rather satisfactorily for the syntactical analysis of a large class of Russian sentences, though its exact reach has not yet been fully determined nor all of its details debugged[25].

The other three subgroups at GU are working on MT as a whole, two of them from Russian into English, the third from French into English. During the last months, the research done at GU has broadened and MT from additional languages into English has begun to be investigated. However, I am not aware of any publications reporting on these new activities and shall therefore not deal with them here. They seem to be at present in their preliminary stages only.

I already mentioned in section 2 that far-reaching claims were made by the GU subgroup headed by Mrs. Ariadne W. Lukjanow and using the so-called <u>Code Matching Technique</u>. I expressed there my conviction that this group could not possibly have developed a method that is as fully-automatic and high-quality as claimed. There are in principle only two procedures by which such claims can be tested. The one consists in having a rather large body of varied material, chosen by some external agency from the field for which these claims are made, processed by the machine and carefully comparing its output with that of a qualified human translator. The other consists in having the whole program presented to the public. None of these procedures has been followed so far. During a recent demonstration mostly material which had been previously lexically abstracted and structurally programmed was translated. When a text lexically abstracted but not structurally programmed was given the machine for translation, the output was far from being high-quality and occasionally not even grammatical. True enough, this did not prevent the reader most of the time from understanding what was going on, but I was told that once or twice the translation was quite wrong, something I could not check personally because of my insufficient knowledge of Russian. In addition, perhaps due to its smallness, the sample did not contain any of those constructions which would cause word-for-word translation to be very unsatisfactory.

The task of evaluating the claims and actual achievements of the Lukjanow subgroup is not made easier by the fact that there seems to exist only one publicly available document prepared by herself[26)]. This document contains 13 pages and is not very revealing. The only peculiarity I could discover lies in the analysis of the source-text in a straight left-to-right fashion, exploiting each word as it comes, including the demands it makes on subsequent words or word blocks, whereas most other techniques of syntactical analysis I know try to isolate certain units first. I shall return to this approach in the next section.

The claim for uniqueness (and adequacy) of the translation of a chemical text is based upon an elaborate classification of all Russian words that occurred in the analyzed corpus with some 300 so-called semantical classes. Though such a detailed classification should indeed be capable of reducing semantic ambiguity I am convinced that no classification will reduce it to zero, as I show in Appendix IV, and that therefore the claim of the Lukjanow group is definitely false. For the benefit of those who need a more palpable refutation, I promise to exhibit a Russian sentence, occurring in a chemical text, which will be either not uniquely translated or else wrongly translated by the Lukjanow procedure, within a week after all the details of this procedure will be in my possession.

On the other hand, I am quite ready to believe that this subgroup has been able to develop valid techniques for a partial mechanization of Russian-to-English high-quality translation of chemical literature (or else for a full mechanization of low-quality translation) -- and this in spite of the poor quality of some publications of other members of this subgroup[27)] -- but, unfortunately, this group seems to be extremely reluctant to make the details of its program publicly available. Should it turn out that they did make some real progress not achieved elsewhere, this reluctance will have caused a great waste of time and money in other MT research groups.

A third subgroup at GU led by Dr. Michael Zarechnak is proceeding in a somewhat different manner, using a so-called General Analysis Technique, and is making less far-reaching claims. Much of its work which I was able to check seemed to me well-founded and to contain solid achievements. However, as this is not the place to go into technical details, it is not possible to present an exact evaluation of where this subgroup stands right now. They hope to be ready with a demonstration within a few months, and I also understand that everybody is welcome to look over their program to the degree that it has already been written up. This group does envisage the utilization of a post-editor for high-quality final output.

With regard to the fourth and last subgroup at GU, led by Dr. A.F.R. Brown, I shall say very little here since I was unable to talk with Brown personally. As already mentioned, he is mostly interested in translation from French into English. I understand from his numerous seminar work papers that he is developing his program on a sentence-after-sentence basis, i.e., dealing with the translation problems as they come and, so I was told, solving them one after another with great ingenuity. I have already expressed my conviction that this approach is wasteful and am sorry indeed that I was unable to talk this over with one of the seemingly most successful adherents of the empirical approach.

Altogether, I think that among themselves the four subgroups at GU do cover pretty well the total realm of problems arising in connection with MT. Dostert's interest in this field stems from his participation in the First MT Conference in June 1952, and so does Garvin's who attended the public opening meeting of this conference. These two linguists have been spending since much of their time on scientific and organizational aspects of MT, and training a large number of other people now working on MT at GU. This is a good deal of experience, and it is therefore not surprising that the work done under their direction should indeed cover, more or less, all the aspects of the MT problem. I am stressing this point since, in spite of the fact that I do disagree with some of the views and approaches of Dostert and his collaborators, I believe that every newcomer to the field -- and there have been many of those during the the last year and more are in prospect -- should make himself as thoroughly acquainted as possible with the work done at GU, and get as clear a picture as possible of their achievements and failures. Otherwise he will have a good chance of repeating work that has been done there, and perhaps repeating the many failures that undoubtedly must have occurred there during the years. There exists no other group in the United States, or in England for that matter, which has been working on such a broad front. This remark of mine is not to be interpreted as implying that the prospective newcomers will not have to get acquainted with anything done outside GU. On the contrary, I do not think that there is much done at GU in the field of MT which is not being done also elsewhere, sometimes in more than one place, and in some of these places, perhaps even more effectively. But GU is still the best place to get a full view of the problem, or rather could be so if each subgroup were equally willing to discuss in full detail its work with others.

10. Having just discussed, far too briefly, the work done by one group in the Washington area, let me now describe, even more briefly, the work done by the other group working on MT in that area, i.e., the group consisting of Dr. Ida Rhodes and one or two associates at the National Bureau of Standards. Dr. Rhodes has been working on this problem for less than a year and there exist no publications so far. It is nevertheless my definite opinion, based upon a few talks during which I was able to go through her program with considerable detail, that her approach is promising and worth close study. Not that she has been able so far to achieve any new results, but I believe that she has been able to obtain old results by sufficiently new and occasionally quite ingenious methods. Dr. Rhodes is one of the few people in the field who has had long experience with actual programming. Being a native Russian speaker, she has been able to combine her linguistic intuitions with her thorough knowledge of computers and their programming into an MT program which, judging from its presently existing outline, should, when fully developed, be able to achieve whatever can be achieved in this field in one of the most efficient and economical ways I am aware of. Dr. Rhodes is a mathematician by training, and her knowledge of modern structural linguistics is very slight. It should furnish some grounds for thought to realize how much of the practical aims of MT can be attained with so little use of structural linguistics. It should, however, be taken into account that Dr. Rhodes' aims are wholly practical, and that no attempt is made by her to obtain a FAHQ output.

Let me mention just one detail in her program. One of the major problems in the syntactical analysis of the given source-language sentence is the problem of where to start. Garvin, for example, instructs the machine to look first for participial constructions and relative clauses. Harris, working with English though, lets the machine look for nominal blocks beginning with the end of the sentence and working backwards. In both approaches, it is of course necessary to go over the sentence a few times before its final analysis is obtained. Dr. Rhodes, perhaps because of her linguistic naiveté, starts the analysis always with the first word of the sentence and lets the machine go over the words one after another, each time rewriting part of its own program, partly recalling Mrs. Lukjanow's technique mentioned above. I do not think that Dr. Rhodes' method in this respect is necessarily better or quicker than the ones adopted, for instance, by Harris and Garvin, but I am also quite sure that it is not necessarily worse. If this is so, then it has certainly the advantage of being transferable in its basic idea to the treatment of the translation from other languages whereas, I presume, Harris' and Garvin's approaches are very much more tailor-made for English and Russian, respectively. In this connection, the interesting question arises, which of these three procedures is closest to the one used by human translators, if human translators use one common procedure at all, which seems to me to be at least highly doubtful. Not that this question is of any practical importance for MT at this moment; however, if and when the time will come when translations will be performed by machines with learning abilities and using, at least partly, rather general heuristic instructions instead of the fully spelled-out program which is customary at present, our question may become a practical one since we would then probably want to give the machine the same or similar heuristic instructions which are given today to human translators during their training or which they develop for themselves in time.

11. Returning from this aside, let us turn now to another of the larger groups engaged in MT research, namely the one at RAND Corporation, Santa Monica,

California, headed by Dr. David G. Hays, with Professor Kenneth E. Harper of UCLA serving as its chief consultant. RAND Corporation has dealt with MT off and on as early as 1950. The well-known study by Professor Abraham Kaplan on the reduction of ambiguity through context[28)] was done at RAND, and Dr. Olaf Helmer of RAND participated in the First MT Conference. However, it is only during the last years that RAND's interest in MT has greatly increased so that the RAND MT group is at the moment one of the large ones. It is there that the empirical approach has found its perhaps strongest expression, probably because Harper is such a strong believer in its soundness. The method they advocate is to go over a certain sample of Russian texts, say of 30,000 words in length, "derive" from a human analysis of this corpus both a dictionary and a set of syntactical and semantical rules, test the derived dictionary and rules on a new sample of the same size, to increase the dictionary and, if necessary, expand as well as improve upon the rules as a result of this test, go on to the next sample, etc. As a matter of fact, during the first six passes -- if I remember correctly, they have just started work on the seventh corpus -- they have mostly tried to perfect the dictionary and solve some of the problems of polysemy -- for example, that bothersome problem of the unique rendering of Russian prepositions. It is only now that they are attacking the question of syntactical analysis. It is impossible to go here into a detailed description of their planned approach but, again, the approach is quite empirical and therefore, in my opinion, wasteful in principle and not too promising in its details as they stand at the moment. This is the more deplorable since Harper is one of the most solid workers in the field of MT with a deep understanding of all its aspects. I already mentioned at the beginning that the report on the state of MT prepared by the RAND group is in general reliable[29)] though I would very often disagree with their evaluation of this state.

Being interested here only in the broad outlines of the state of MT, I am not going to mention all the very many specific contributions made to the treatment of MT in its innumerable aspects by the RAND group or any other group mentioned so far or to be mentioned later on. Some of these contributions are independent of the general attitude since they may be dealing with such questions as the most efficient method for transliteration, questions of coding, instructions for the keypunch operator, etc., all of which are important aspects of any practical MT procedure, though I myself shall not discuss them any further in this report.

It is interesting, in view of some remarks I made above, that the RAND group intends to deal with the problem of syntactical ambiguity by taking into consideration those words in the sentence which are immediately contiguous to the one whose syntactical status is ambiguous in isolation. This procedure is, of course, rather natural and consciously, or unconsciously, based upon the immediate constituent model discussed above. Knowing that this model is not a fully adequate one, I am not impressed by the claim that resolution of syntactical ambiguities by consideration of the immediate neighborhoods of the ambiguous expression has proved itself in practice. Let me state, however, for the sake of fairness, that a report which is probably inspired by Harper, if not actually written by him[30)], contains a statement to the effect that its author is not very much impressed by the fact that counter-examples of his empirically derived rules can be concocted so long as these are concocted examples and not ones that occur in some actual text. Final judgment of this issue must be left to the reader.

12. In the area of Greater Los Angeles there is another group working on MT at Ramo-Wooldridge Corporation, headed by Dr. Don R. Swanson. Harper worked as a consultant for this group at an earlier stage, and there exists a close cooperation between the RAND group and the Ramo-Wooldridge one. Though there

are some differences in their approach, a description of these differences would require going into greater detail than I am prepared to do here. This is the group which published the interesting report mentioned in the preceding paragraph, as well as an even more interesting and very detailed report on the latest phase of its activities[31)], a close study of both of which I would suggest to everybody in the field. Work on MT at Ramo-Wooldridge has also had its ons and offs, according to the amount of contract money available, and the financial future of the group seems to be unsettled. I hope, however, that Swanson at least will be able to continue his work on MT in some form or another, since he has been able to make some solid contributions to the field in the past and will doubtless be able to do so in the future.

13. Another small group whose philosophy is closely related to that of RAND and Ramo-Wooldridge is the one working in the Willow Run Laboratories of the University of Michigan, Ann Arbor, headed by Mr. A. Kotsoudas. There is nothing I can say about the activites of this group beyond the statement contained in the NSF booklet, except that it does not seem to me that this group has made any specific lasting contribution to MT so far.

14. The last of the older groups, i.e., those in existence for more than a year, is the Harvard University group headed by Professor Anthony G. Oettinger. It is quite amazing to find that this group still busies itself almost exclusively with an exploration of the word-by-word translation method. There seems to exist in it a strong distrust of the achievements of other groups. Though it may well be admitted that the possibilities of a word-by-word translation from Russian into English have never before been so thoroughly explored as they were by this group, with some new insights gained, and that very valuable results were obtained relative to the structure of MT dictionaries, one still wonders whether

the right proportion between utilizing other people's work in the field and distrusting their work has been struck by this group -- though there certainly are good reasons for the distrust on quite a few occasions. Since there exists an extremely detailed and easily accessible account of its work[32)], I shall here say no more about it.

15. This leaves us with two more American groups that started their work very recently. There is one at Wayne University, Detroit, headed by Professor Harry H. Josselson and Dr. Arvid W. Jackobson, a linguist and a computer mathematician, respectively, and the one at the University of California, Berkeley, directed by Dr. Louis Henyey and Dr. Sydney Lamb. No results have been achieved by these groups so far nor, of course, could they have been expected in this short time. Let me make only a few short comments on their methodology.

The Wayne group expects to deal with the same problems treated elsewhere, but intends to make more use of modern statistical techniques. I am not quite sure what exactly this is supposed to mean, and I have already expressed my doubts as to the effectiveness of the analyzing of a huge corpus of text, to which alone statistical methods would be applicable, for MT purposes beyond certain obvious points.

The Berkeley group, on the other hand, having originally advocated also a strongly "empirically" directed approach seems to have changed its mind somewhat and is now trying to strike a middle way between the divergent philosophies. Though I believe that this group did not yet find an optimum compromise, its program strikes me as quite reasonable and promising if only it will be able to avoid the ever existing danger of just more or less repeating, perhaps a little better, what other people have been doing before.

16. There exists another small group at the University of Texas, Austin,

headed by Professor Winfred P. Lehmann. I know nothing about the activities of this group, except that it must have started its work rather recently, and that judging from a talk given by Lehmann at the GU Round Table Meeting on MT[33], they are working on German syntax.

17. He who has read, or will read, the report of Reitwiesner-Weik (R-W) mentioned in section 1, will notice that the three groups discussed in the last two sections have not been mentioned by them. This is simply due to the fact that these groups started their work after the completion of the R-W report. On the other hand R-W do mention activities which have not been treated by me. An explanation is in order.

The National Science Foundation (R-W, pp. 11-12) is sponsoring research on MT and organizing conferences on MT but is not conducting research on its own. In addition, it is publishing very valuable reports on Current Research and Development in Scientific Documentation, part of which is dedicated to MT. The latest of these reports[34] deals with MT on pp. 31, 32, 38, 39, 42, 47-57.

The U.S. Air Force, Air Research and Development Command (R-W, p. 12) sponsors and supervises MT research but does not seem to be engaged itself in it. The final report of the University of Washington group mentioned there as due for about May 1958, has not yet appeared but should be out very soon.

The U.S. Army (R-W, p. 13) only supports research.

The research at Bell Telephone Laboratories and Haskins Laboratories (R-W, pp. 14-15) is only very remotely related to MT.

IBM (R-W, p. 14) joined forces in 1953/54 with the GU group in the preparation of the well-known GU demonstration. Dr. Gilbert W. King, formerly with Telemeter-Magnetics, Inc., Los Angeles, California, joined IBM in 1957 and is in the process of organizing an MT group at the IBM research center in

Yorktown, N.Y.

Research at Telemeter-Magnetics, Inc., itself (R-W p. 17) has been discontinued, to my knowledge, after King left. Incidentally, the name given to the high-capacity photoscopic storage disc developed by King, "The USAF Automatic Language Translator Mark I", is highly inappropriate.

Research at the California Institute of Technology, Pasadena (R-W, pp. 17-18) was also discontinued after Mr. Toma left in 1958 in order to join the GU group. He is now working with the subgroup headed by Zarechnak.

The research at Indiana University, Bloomington (R-W, p. 20) is only remotely relevant to MT in its present stages.

I am not aware of any recent activities in MT at the State College of Washington, Pullman (R-W, p. 22).

The University of California at Los Angeles (R-W, pp. 22-23) was one of the first centers of MT -- we recall that Professors Kaplan, Victor A. Oswald, Jr., William E. Bull and Harper were, and still are, teaching there -- but to my knowledge none of them is now working on MT at the University, though Harper is serving at the moment as a full-time consultant to RAND Corporation and the others might still occasionally do some consulting on MT matters, too. Oswald gave a talk in the GU Round Table Meeting in 1957[35].

I do not think that the work at the University of Chicago (R-W, p. 23) is of any specific relevance to MT.

Equally irrelevant is the work done at Western Reserve University, Cleveland, Ohio (R-W, pp. 30-31).

18. Before I go on to describe the work of the two British groups, let me stop and try to summarize the situation as it exists at present in the United States. Most groups dedicate most, if not all, of their work to Russian-to-English translation. The only exceptions are the MIT group which works, in so much as

it works on straight MT problems at all, on German-to-English translation, the Seattle group which spends part of its time on German-to-English translation and apparently a little also on other pairs of languages, and one subgroup at GU which works on French-to-English translation. The Philadelphia group is exclusively concerned with the syntactical analysis of English. The concern with Russian as a source-language is, of course, no accident, but due to the simple and well-known fact that translation of scientific, technological and intelligence material from this language is of vital concern to American science, technology and security. It is probably also no accident that those MT groups aiming at short range results, and therefore willing to renounce, at least for the near future, the ideal of FAHQMT have concentrated on Russian-to-English translation whereas the more theoretically minded groups were working mostly with other pairs of languages.

There is little I would like to say at this point on the prospects of the theoretical approach in addition to what I have already said before. Research along this line should definitely be treated as a long-range, highly basic activity whose direct practical applicability at any time is rather doubtful though, as said before, it might lead to important insights into the workings of language, and therefore sooner or later also to some indirect practical applications.

With regard to that part of MT research which is oriented towards achieving practical results in the near future, let me make the following comments. The whole gamut of problems of Russian-to-English translation is covered at present (with the obvious exception of preparing additional idioglossaries). I already said that at GU alone all or most of these problems are treated. Hence, in regard at least to the topic, if perhaps not so much with regard to the method, there must exist a considerable overlap between the activites of the various groups. It is my definite opinion -- which it would be extremely difficult if not impossible to document, and certainly not within the limits of this report -- that this overlap is too high, and that at least fifty percent of the current research effort is wasted in the sense that either known results are obtained anew by the same or not significantly different methods, or that old failures are repeated. I would not want to deny that in a certain sense both rediscovery of old results and repetition of old failures may have their value. But I would still say that these advantages could and should be obtained more cheaply and less wastefully.

There does exist some cooperation between certain groups, especially the smaller ones, but it is also a well-known fact that some groups are quite reluctant to share their detailed results with others, perhaps because of a feeling that these results have not yet gotten their definitive formulation, perhaps also for less altruistic reasons; in this connection I think I should mention especially the Lukjanow subgroup at GU and the MIT group. On the other hand, there are groups which feel that they have little to learn from other groups' achievements and, if I am not mistaken, the Harvard group is a good example of this attitude. I would guess that if nothing is done to improve this state of affairs, not only will valuable research money be wasted, but the actual going into business of a man-machine partnership in Russian-to-English translation might be postponed beyond necessity for a couple of years or so.

The need for constant and more elaborate exchange of ideas has been repeatedly expressed by the leaders of many MT groups with the NSF explicitly offering its help in this respect, but it seems that so far no really effective measures have been taken to put this collaboration into practice. I shall, at the end of my report, make some definite recommendations in this direction.

19. Let me now turn to the two groups working on MT in England. One of these is operating at Birkbeck College in London, headed by Dr. Andrew D. Booth, the other at the University of Cambridge, headed by Mrs. Margaret Masterman-

Braithwaite[36]. Booth is one of the very first persons who thought of the utilization of electronic computers for translation as early as 1946, and has written together with Dr. R.H. Richens, presently a member of the Cambridge group, a pioneer paper on MT in 1948[37]. He has continued his research in this field almost uninterruptedly though always only part-time and published last year, together with two associates, a book dealing mostly with machine translation[38]. He was also one of the editors of the first book dealing with MT[39], which contained 14 monographic studies on various aspects of the MT problem, in addition to a foreword by Dr. Warren Weaver of the Rockefeller Foundation and a valuable historical introduction. His recent book contains a great wealth of insights into the syntactical structure of German, and to a lesser degree into that of French and Russian, but the approach suffers from an excessive adherence to the empirical method in so much as rules for resolving syntactical ambiguity are based, in principle, "on analysis of all the existing literature on the subject in question", and in practice, for the purposes of illustration, on the analysis of a very small amount of text. The same holds for the methods proposed in this book for the reduction of semantical ambiguities. The authors are aware of the limitations of this method but intend to leave the development of a method that would resolve ambiguities in all conceivable (scientific) texts to people with a high degree of acquaintance with the German language. Many statements, of either historical or systematic nature, made in this book are sometimes rather cavalier, and could create a somewhat distorted picture, especially with regard to the relative importance of the insights gained by this group itself. There is, however, no point of here going into such details. The book contains, in addition, many technical details of the construction of programs for MT, a full account of which may be gained from a companion volume by Booth's wife[40].

20. It might be worthwhile mentioning that this book also contains a refutation of one very frequent argument for the use of an interlingua, i.e. an artificial mediating language, for MT purposes[41]. This argument points out that translation from each of n natural languages into each other requires the establishment of $n(n-1)$ programs (including dictionaries and idioglossaries) whereas the use of an interlingua, into which and from which all translation exclusively proceeds, requires only $2n$ such programs. (For ten languages, for example this means a reduction from 90 to 20 programs.) The fallaciousness of this argument is immediately obvious, however, as soon as one realizes that using one, any one, of the original n languages as a mediating language would reduce the number of programs even more, namely to $2(n-1)$ (in our illustration to 18). This counter-argument does not, of course, prove that the idea of using an artificial mediating translation language is wrong as such, and other arguments have been brought forward in its defense, but the one refuted just now seems to have been one of the most potent ones, and with its elimination proponents of the interlingua idea should give it a second thought.

It should indeed be carefully tested, for independent reasons, to what degree the quality of a translation between two languages is impaired, if instead of a direct translation, an indirect one is employed, based upon high-quality translation from the source-language into some intermediate language and from it into the target-language. So far there exist, to my knowledge, only more or less anecdotal results in this respect. Should it turn out that high-quality translation is generally obtainable by going through some intermediate language, natural or artificial, this would be of enormous importance for multilingual MT of the future.

Whereas the mentioned argument "from n^2 to $2n$" for the use of an artificial interlingua in MT can definitely be proven fallacious, though it holds good as an argument for the use of any intermediate language, there are of course other arguments to support the use of an artificial interlingua qua artificial, whether of the Esperanto type or of that of a symbolic language system. I admit that the idea of a "logical", unambiguous (in every respect, morphologically, syntactically and semantically) interlingua has its appeal today as had the related idea of a characteristica universalis in the 17th and early 18th centuries. This appeal is bolstered by the great achievements of modern mathematical logic with its constant use of artificial language systems, and there is therefore some force in the claim that an idea that failed in the 17th century need not do so in the 20th. But the present argument is no less fallacious. Its fallacy lies in the assumption that "translation" from a natural into a "logical" language is somehow simpler than translation from one natural language into another. This assumption, however, is totally unwarranted, whatever its appeal to someone with little direct experience with symbolic language systems. As a matter of fact, the transition from a sentence in a natural language to its counter-part in a language system deserves the name 'translation' only in a somewhat Pickwickian sense. I shall not elaborate this point any further, but only mention that it has been discussed rather widely in recent methodological literature. We have here probably another result of the customary loose use of the word 'translation' which has already caused a lot of trouble on other occasions (such as in connection with information retrieval where the issue becomes constantly befuddled through an uncritical and still more metaphorical use of this word). Not only is the process of presenting a counterpart of some natural language sentence in some symbolic language system in general incomparably more difficult than its translation into some other natural language even for a human being, as everyone who has ever taught a freshman course in symbolic logic will readily certify, but the mechanization of this kind of 'translation' poses problems which are by orders of magnitude more difficult than those posed by translation proper. It is no accident, again, that not only have linguists not attacked these problems in any serious sense, but that even hard-boiled logicians have shunned it in favor of dealing with "easier" ones (which ordinary linguists regard as lying beyond their comprehension). Altogether, the problems revolving around an interlingua as a device for MT are still in a highly speculative state, and it is probable that years will pass before any practical results can be expected.

21. This brings us to the second British group in which the idea of an interlingua has played a decisive role in the latest aspects of its ways of thinking[42]. In spite of its constant disclaimers, I regard this group as a highly speculative one with many of the good and equally many of the bad connotations of the term. I find myself again and again amazed by the prolificy of ideas emerging from this group, almost all of which have some initial appeal while also having the disturbing property of constantly changing their exact meaning or being quickly replaced by some other idea, for which the same process starts all over again after a very short time. I myself, in the early stages of my thinking on MT, have played with many of these ideas and can therefore readily testify to their appeal. I did, for instance, repeatedly spend some time on the question of whether and to what degree Combinatory Logic could be applied to MT, and though I have failed so far to achieve any serious results in this connection, I am not convinced that I myself, or other people better equipped for this purpose, could not still do so if working very hard and uninterruptedly on this problem. In one of my publications I made a brief mention of this issue[43]. Miss Masterman wrote three years ago a long (unpublished) paper on this topic, but I had great trouble understanding its point, and the issue is no longer mentioned in more recent publications of the Cambridge group, having apparently been superseded by the idea of applying lattice theory[44]. Now lattice theory is the theory of a structure which is

so general that one should not be surprised to find it embodied in some actual situation. There can also be no doubt that lattice theory, and certain more general branches of abstract algebra such as the theory of semi-lattices and partially ordered systems, can be applied to linguistic investigations though I am not aware of any new insights gained so far by such applications. The applications made by the Cambridge group of their lattice-theoretical approach, inasmuch as they are valid, are only reformulations in a different symbolism of things that were said and done many times before.

A third idea emerging from this group, though not only from it, is that of using a thesaurus-type dictionary instead, or perhaps in addition to, ordinary dictionaries. I find here the greatest difficulties of understanding in spite of many attempts on my part to do so and many hours of talking with various members of the group. Among other troubles I have here is the fact that the term "thesaurus" has not only been used by various groups in different, occasionally quite different, senses, but that members of the same group often use the term in different senses, and that its meaning keeps shifting even in the publications of one and the same person with no adequate warning given to the reader, perhaps without the writer being aware of such a shift. So we find that a thesaurus is sometimes meant to be rather similar to Roget's well-known Thesaurus of the English Language, and sometimes explicitly rather different from it, in which case not always an indication is given of the specific character of the intended difference. Sometimes the thesaurus is supposed to contain after each entry so many expressions of the same language; sometimes it is supposed to contain, perhaps in some code, the interlingua equivalents of these entries, and so on[45].

The only sound idea I can see behind all this fuss about the thesaurus is the old idea already expressed in Weaver's memorandum of 1949 that the ambiguity of words of the source language in isolation is reducible through taking proper account of its linguistic environment. One has tried many times to write a program exploiting this idea, but so far never with full success. The main trouble is that the word, or the words, which can serve as clues for this reduction of ambiguity do no always occur in the immediate neighborhood of the ambiguous word, say one or two words on either side of it, though this will happen most of the time. Sometimes not even a whole sentence,or a whole paragraph, for that matter, would be a sufficiently large environment for complete reduction of ambiguity by machine though it might be so for an intelligent human reader. A demonstration of this contention of mine is given in Appendix IV.

Altogether there exists so far no evidence that any of the ideas brought forward by the various members of the Cambridge group will ever contribute new effective methods for practical MT, and little evidence that they would result in new valid insights into the workings of language.

22. I understand that the National Physical Laboratory, Teddington, England, is in the process of organizing a group that will work on MT. So far, however, I know of nothing more specific in this respect. I am not aware of any organized research on MT outside of the United States, England, Russia and Italy, where Dr. Silvio Ceccato heads a small research group at the University of Milan. Since this group apparently has not yet published its recent findings and since I could not persuade myself that I understood the articles published by Ceccato three years ago[46], I shall say no more about this last group here. (My own recently created, small group at the Hebrew University, Jerusalem, Israel, has so far done very little constructive work on MT.)

A fully detailed, critical report of the state of research on MT in the USSR is highly desirable. I am in no position to present such a report myself, not having visited there. I understand, however, that Oettinger, who had an opportunity to visit three of the five (at least) Russian research centers during the summer of 1958, has prepared a short report which will be published soon. Some equally recent impressions on the state of MT in Russia can be gotten from reports prepared by Professor John W. Carr, III, and Professor Alan J. Perlis who, during the summer of 1958, visited various computer installations in Russia, and had an opportunity to talk briefly with representatives of most MT groups[47)]. However, since MT was not their major concern, the picture one gets from these reports is not as detailed and critical as one could wish.

23. Let me summarize and make some proposals:

(1) Fully-automatic, high-quality translation is unattainable in the near future, and not attainable altogether unless machines can be built and programs for them written which will endow these machines with quasi-human intelligence, knowledge and knowledgability.

(2) Basic linguistic research is of great importance as such, and should be supported whether or not it will lead to improvements of MT techniques. Most of this research would gain if applications to MT problems will not be taken into account from the beginning.

(3) For the time being, research on MT proper should only concern itself with supplying mechanical aids to translation, while aiming at constantly improving these aids and increasing their number. By pooling the available, highly dispersed knowledge in the field, it should be possible to establish within a period of a few years translation centers that would be able to compete commercially with existing all-human translation establishments either in providing high-quality translations while requiring only a fraction of the human time invested there, thereby making a substantial contribution to the practical translation problem, or, alternatively, in producing low-quality

translations without any human intervention.

(4) The economic basis of a commercial partly-mechanized translation center would be strengthened by the development of a reliable print reader and the construction of a special-purpose translation machine. These two developments should therefore be given high priority.

(5) The special-purpose translation machine should be constructed in such a way that it could conveniently be programmed to come up with an output that would enable a human post-editor to produce a high-quality translation, as well as with an output that could stand by itself as a low-quality translation surrogate that would be satisfactory in those situations where no more is required.

(6) The damage done by a failure to pool available knowledge and to plan an efficient division of labor will probably be much greater than the advantages gained from a spirit of competition, and might cause a delay of two or three years in the establishment of a working, commercial, partly-mechanized translation center. It probably makes little difference whether one large research and development center for MT is created or whether the existing groups come to an agreement as to their pooling of knowledge and division of labor.

(7) Since most, if not all, of the research funds are supplied from government and military agencies, it should be not too difficult to obtain a degree of cooperation which would insure full utilization of the achievements attained so far for the purpose of quickening the pace towards the establishment of the first commercial translation center.

(8) Not only should the existing research staff be encouraged to cooperate and be given the opportunity of quickly exchanging ideas, half-baked or fully-baked, but more people should be trained to deal with the countless still unsolved MT problems, especially, of course, with regard to translation

into English from other languages than Russian and from English into various languages.

(9) I propose, more specifically, that during the summer of 1959 various high level seminars should be organized by some such organization as the National Science Foundation, in which one or two leading members of each of the groups engaged in practical MT research should exchange their knowledge on selected specific aspects of MT, thrash out their differences and arrive, if possible -- and I think it should be possible --, at the determination of one or two most promising methods for the solution of each such specific problem. Subsequently they should come to an agreement as to which group or groups should undertake the detailed solution of these problems. There are groups who claim to be in possession of more or less complete solutions of certain specific problems. These claims should be checked, preferably with the help of a computer that, together with a staff of programmers, should be made available for these seminars. If the claims are sustained, the respective problems could safely be regarded as solved though there will, of course, always be room for still better methods. These seminars should each last for about a month.

(10) Starting in September 1959 or, should it be too late to arrange for this, in September 1960, one or two universities or technological institutes should undertake the training of suitable candidates for work on MT, whether as research workers, programmers or post-editors. Georgetown University, MIT-Harvard University, the University of California (Los Angeles and/or Berkeley), and the University of Michigan are the most likely places where such instruction could be given. This would be a one year course for students with B.A. or B.Sc. degrees at least. The exact curriculum of such a course could probably be worked out without too much difficulty.

24. Let me wind up with a remark on bibliography. I did not deem it

worthwhile to provide one, in addition to the literature mentioned in the notes, since there exist already quite a number of such bibliographies and a bibliography of Russian publications is in preparation, as mentioned in section 21. The most important one is, of course, the annotated bibliography appended to each issue of the journal _MT_. The last issued in my possession, Vol. 4, No. 3, is dated December 1957 but must have appeared at the earliest in March 1958, since one of the reports mentioned in the bibliography appended to this issue is dated February 1958. The last item in the bibliography of this issue carries the ordinal number 134. I understand that Vol. 5, No. 1 appeared in December 1958, but this issue has not yet come into my possession. Many of the items mentioned in my notes are not yet contained in this bibliography.

Other bibliographies are given on pp. 227-236 of _op. cit._ in note 14 (46 annotated items), pp. 82-95 of _op. cit._ in note 10 (82 annotated items), pp. 20-51 of _op. cit._ in note 4, and pp. 51-65 of _op. cit._ in note 5 (containing some 170 items, including internal reports, work papers, etc.). Useful current references ar given _passim_ in _op. cit._ in note 6.

It would be very helpful, if someone, perhaps the editors of _MT_, would publish a consolidated, annotated bibliography covering the first decade of MT, 1949-1958.

NOTES

1) This estimate is not official. In addition, it is still rather difficult to evaluate available machine time. Some basis for the estimate is provided in Appendix I.

2) Reitwiesner and Weik, in their report mentioned below in note 5, say on page 34 that "Dr. Panov's group consists of approximately 500 mathematicians, linguists and clerical personnel, all working on machine translations of foreign languages into Russian and translations betwwen foreign languages with Russian as an inter-language".

3) Y. Bar-Hillel, "The present state of research on mechanical translation," American Documentation 2:229-237 (1951, appeared 1953).

4) H.P. Edmundson, K.E. Harper and D.G. Hays, "Studies in machine translation -- 1: Survey and critique", Project RAND Research Memorandum RM-2063, February 25, 1958. Eight more memoranda were published in this series in December 1957 and during 1958.

5) G.W. Reitwiesner and M.H. Weik, "Survey of the field of mechanical translation of languages", Ballistic Research Laboratories Memorandum Report No. 1147, May, 1958.

6) "Current research and development in scientific documentation, No. 3", NSF-58-33, Science Information Service, National Science Foundation, October 1958.

7) "Research in machine translation", Monograph Series on Languages and Linguistics No. 10, Georgetown University Press, Washington, D.C., 1957.

8) A.D. Booth, L. Brandwood and J.P. Cleave, Mechanical resolution of linguistic problems, Academic Press Inc., New York and Butterworth Scientific Publications, London, 1958.

9) I have to beg the reader's pardon for this seeming pleonasm. But 'machine translation' has apparently come to mean translation-with-some-use-of-machinery so that it is not really pleaonastic to speak of 'fully-automatic machine translation' nor contradictory to speak of 'partially-automatic machine translation'.

10) This estimate is given on p. 58 of "Design Study for an Intergrated USAF Intelligence Data Handling System, Appendix A, Machine Translation of Languages",

submitted by the Ramo-Wooldridge Corporation, 31 March 1957.

11) In addition, whereas the estimate of one fourth of a cent was based on a rate of 20 Russian words per minute (ibid.), in the RAND report mentioned in note 4, p. 12, the maximum rate of trained and experienced keypunch operators is given as 600 words per hour. This alone doubles the expense.

12) On p. 57 of the report mentioned in note 10, it is estimated that an automatic print reader might be ten times cheaper than human retyping. This estimate is doubtless highly speculative. It is strange that the estimates on human translation cost diverge so greatly. The latest estimate I know of is given as "1 to 3 cents per word" on p. 5 of "Experimental machine translation of Russian to English", Ramo-Wooldridge Project Progress Report M20-8U13, 15 December 1958. I have already heard mentioned the figure "4 cents per word" and even higher ones. These figures need not be commensurable as the specific form of the final human output is usually not given.

I understand that a certain outfit in Israel which does large-scale translation of scientific material from Russian to English for an American agency charges about 2 cents per word for a finished product.

13) So long as keypunching will be used for the input, it will doubtless be highly profitable to introduce as much pre-editing as the keypunch operator can take into stride without slowing down to any considerable degree. The problem will become more delicate after a print-reader takes over.

14) These are the terms explicitly introduced for MT purposes on p. 88 of W.E. Bull, Ch. Africa and D. Teichroew, "Some problems of the 'word'", Machine Translation of Languages (W.N. Locke and A.D. Booth, eds.), Technology Press of MIT and John Wiley & Sons, New York, Chapman & Hall, London, 1955, pp. 86-103.

15) This term has already caused a lot of confusion. Cf., e.g., p. 172 of the book mentioned in note 7. In this report, however, its meaning should be unambiguously clear from the following sentences. The clearest presentation of this

approach is given in H.P. Edmundson and D.G. Hays, "Studies in machine translation - 2: Research methodology", Project RAND Research Memorandum RM-2060, December 16, 1957. Cf. note 4.

16) This talk is reproduced on pp. 514-518 of the Proceedings of the Eighth International Congress of Linguists, Oslo University Press, Oslo, 1958. The reports and discussions of the section meeting on MT, which I chaired, are reproduced on pp. 502-539.

17) For this topic, see E. Reifler, "Mechanical determination of the constituents of German substantive compounds", MT 2:3-14 (1955). In the Oslo talk (see previous note), Reifler made some very far-reaching claims in this respect which sounded hardly believable. I hope that the promised report will allow for a test of these claims.

18) See p. 577 of the book mentioned in note 16.

19) The language in which Yngve puts his beliefs is rather indefinite. In one of his last publications, "The feasibility of machine searching of English texts", to appear in the Proceedings of the International Conference for Scientific Information, Washington, D.C., November 1958, he says, for instance: "It is the belief of some in the field of MT that it will eventually be possible to design routines for translating mechanically from one language to another without human intervention" (p. 167 of the preprinted volume, Area 5). It is rather obvious from the context that Yngve includes himself among the "some". How remote "eventually" and "ultimately" -- another qualifying adverb occurring in a similar context -- are estimated to be is not indicated.

20) Among Chomsky's many pertinent publications, I shall mention here only his book Syntactic Structures, Mouton & Co., 's-Gravenhage, 1957, which also contains further references.

21) This talk, revised, was published in Logique et Analyse, Vol. 2, No. 5, January 1959.

22) In my paper, "A quasi-arithmetical notation for syntactic description", Language 29:47-58 (1953).

23) This language, called COMIT, is described in an internal memorandum, "A programming language for mechanical translation", dated September 2, 1958. I understand from correspondence with Yngve that other groups are planning to apply COMIT to their own research.

24) "Linguistic transformations for information retrieval", to appear in the Proceedings. It is preprinted on pp. 123-136 of Area 5.

25) Among the numerous relevant publications of Garvin, let me mention only various Seminar Work Papers of the Machine Translation Project of Georgetown University (the latest of which, MT-73, was published in 1958), his contribution, "Linguistic analysis and translation analysis", to the monograph mentioned in note 7, and "Syntactic units and operations" on pp. 626-632 of the book mentioned in note 16.

26) "Statement of proposed method for mechanical translation", Seminar Work Paper MT-35 of the Machine Translation Project of Georgetown University, 1957.

27) It is impossible not to react to Dr. William M. Austin's paper "Language as symbolic logic", op. cit. in note 7, pp. 39-43. Dr. Austin may be a good linguist, for all I know, but what compels him to exhibit in public his total confusion in matters of symbolic logic?

28) "An experimental study of ambiguity and context", The RAND Corporation, P-187, November 30, 1950; published in MT 2:39-46 (1955).

29) Among the exceptions should be mentioned the characterization (on p. 14 of op. cit. in note 4) of the Polish logician Ajdukiewicz as a linguist and the similar mistake with regard to the Polish school of logicians. I myself am characterized on this occasion as the exponent of the Polish school in the United States, which is misleading in various ways. (It is true, however, that I acknowledged in my paper mentioned in note 22 the impact of a certain paper of Ajdukiewicz's which does not seem to have been read by the RAND group, though it appears in their

bibliography.)

30) This is the report mentioned in note 10. The paraphrased statement occurs on p. 39.

31) This is the report mentioned in note 12.

32) A.G. Oettinger, W. Foust, V. Giuliano, K. Magassy, and L. Matejka, "Linguistic and machine methods for compiling and updating the Harvard Automatic Dictionary", to appear in the Proceedings mentioned in note 19. The preprinted version is on pp. 137-159 of Area 5.

33) "Structure of noun phrases in German", op. cit. in note 7, pp. 125-133.

34) See note 6.

35) "The rationale of the idioglossary technique", op. cit. in note 7, pp. 63-69.

36) It is perhaps not superfluous to point out, in view of such descriptions as given in R-W, p. 35, that Miss Margaret Masterman and Mrs. M. Braithwaite and even Miss Masterson (!) in Oswald's paper mentioned in note 33, are one and the same person. Miss Masterman is married to Professor Richard B. Braithwaite of the University of Cambridge, England.

37) This paper, "Some methods of mechanized translation", was reproduced in mimeograph for the First MT Conference in June 1952.

38) See note 8.

39) See note 14.

40) K.H.V. Booth, Programming for an automatic digital calculator, Academic Press Inc., New York, Butterworths Scientific Publications, London, 1958.

41) See op. cit. in note 8, p. 293.

42) The latest version of this idea is described in R.H. Richens, "Tigris and Euphrates -- a comparison between human and machine translation", Paper 2-4 presented at the Symposium on the Mechanization of Human Thought, held at the National Physical Laboratory, Teddington, Middlesex, England, November 24-27, 1958. The paper and subsequent discussions will be published in the forthcoming Proceedings

of this Symposium. Cf. also M. Masterman, R.M. Needham, and K. Spärck Jones, "The analogy between mechanical translation and library retrieval", to appear in the Proceedings mentioned in note 19. The preprinted version is on pp. 103-121 of Area 5. Both papers contain further references. A strong, though by no means conclusive, case for Interlingua (with a capital 'I') as an interlingua for MT is made by A. Gode, "Signal system in Interlingua", MT 2:55-60 (1955).

43) On p. 55 of op. cit. in note 22.

44) See M. Masterman, "New techniques for analyzing sentence patterns" (Abstract), MT 3:4-5 (1956).

45) Among the numerous publications using the term 'thesaurus' in connection with MT, let me mention only the following: M.A.K. Halliday, "The linguistic basis of a mechanical thesaurus", MT 3:81-88 (1956) (cf. also pp. 527-533 of op. cit. in note 16), M.M. Masterman, "The thesaurus in syntax and semantics", MT 4:35-43 (1957), the second of the papers mentioned in note 42, a report for NSF by Gilbert W. King on the work of the Cambridge Language Research Unit, "A thesaurus-lattice approach to the structure of language and communication with words", July 1958 (which made no more sense to me than the publications of this Unit themselves). Curiously enough, a thesaurus approach is also adopted by David G. Hays, "A projected study of semantic ambiguity", RAND Corporation P-944 A, September 24, 1956. I have not heard since of this projected study.

46) The more important one is "La grammatica insegnata alle machine", Civiltà dalle Machine, Nos. 1 and 2, 1956.

47) I have in my possession a mimeographed "Report on a return visit to the Soviet Union by four American digital computer specialists", Department of Mathematics, University of Michigan, Ann Arbor, November 8, 1958, by John W. Carr, III. I attended Perlis' oral presentation of his report in November 1958 but have no copy of it.

If someone doubts that all the aspects of MT covered in my report in connection with the state of the art in USA and England are also fully covered in the USSR, let him look at the "Abstracts of the Conference on Machine Translation (May 15-21, 1958)" a translation by the U.S. Joint Publications Research Service, dated 22 July 1958, from the Russian original published by the First Moscow State Pedagogical Institute of Foreign Languages under whose auspices this conference convened. This brochure contains abstracts of 71 papers that were read on this occasion.

As an interesting sidelight on human translation, let me mention that V.V. Ivanov's paper on "Gödel's theorem and linguistic paradoxes" (p. 20) was rendered by the translator as "Hegel's theorem..." thereby causing me (as well as Gödel himself and other logicians to whom I told the story) a good deal of amusement and -- headache (until the mistake was discovered). Who the man "Lotze" could be who -- according to the abstract -- generalized "Hegel's theorem", I still do not know. (There was a German logician by this name at the end of the 19th century.) The reader will have some fun in trying to reconstruct this comedy of errors.

Another very disturbing error occurs on p. 6 (and elsewhere). I again pondered for hours to find out what could possibly be meant by "methods used in theory of numbers applied to investigation of the grammatical structure of language". I found out at last: "theory of numbers" is a mistranslation for "theory of sets". (Readers who know Russian will easily understand the rationale of this mistake.) It is quite clear from the translation that the translator, while knowing Russian very well and probably being a native Russian, knows English somewhat less and has very little knowledge of modern logic and mathematics.

APPENDIX I

MT Statistics as of January 1, 1959

(No responsibility as to the accuracy of the figures is undertaken. They were obtained by personal communication, the author's impressions or <u>bona fide</u> guesses. In cases of pure guesses, a question-mark is appended.)

Institutions	Year of start of research	Number of workers	Full-time equivalents	Current yearly budget in dollars	Project leader(s)
Georgetown University The Institute of Languages and Linguistics Machine Translation Project 1715 Massachusetts Avenue Washington, D.C.	1952	30 ?	15 ?	?	Leon E. Dostert Paul L. Garvin Ariadne W. Lukjanow Michael Zarechnak A.F.R. Brown
The RAND Corporation 1700 Main Street Santa Monica, California	(1950) 1957	15	9	?	David G. Hays Kenneth E. Harper
Harvard University The Computation Laboratory Machine Translation Project Cambridge 38, Massachusetts	1953	11	7 ?	?	Anthony G. Oettinger
Massachusetts Institute of Technology Research Laboratory of Electronics and Department of Modern Languages Cambridge 39, Massachusetts	1951	10 ?	6 ?	?	Victor H. Yngve
University of Washington Department of Far Eastern and Slavic Languages and Literature Seattle, Washington	1949	10 ?	6 ?	?	Erwin Reifler
University of Michigan Willow Run Laboratories Ann Arbor, Michigan	1955	10 ?	6 ?	?	Andreas Kotsoudas

Wayne State University Department of Slavic Languages and Computation Laboratory Detroit, Michigan	1953	10	6	40,000	Harry H. Josselson Arvid W. Jackobson
University of California Computer Center Berkeley, California	1958	8	5	40,500	Louis G. Henyey Sydney M. Lamb
National Bureau of Standards Washington, D.C.	1958	3	2	25,000	Ida Rhodes
University of Pennsylvania Department of Linguistics Philadelphia, Pennsylvania	1956 ?	10 ?	3 ?	?	Zellig S. Harris
University of Texas Department of Germanic Languages Austin 12, Texas	1958	?	?	?	Winfred P. Lehmann
Various other American institutions and individuals		50 ?	10 ?	?	
Total, USA		150 ?	80 ?	$1,500,000	
Birkbeck College Department of Numerical Information London, England	(1947) 1955	6 ?	3 ?	?	Andrew D. Booth
Cambridge Language Research Group Cambridge, England	1955 ?	20 ?	5 ?	?	Margaret Masterman
Hebrew University Jerusalem, Israel	1958	5	1	4,000	Yehoshua Bar-Hillel
University of Milan	1958	?	?	?	Silvio Ceccato

Other groups and individuals outside Russia		10 ?	3 ?	?
USSR	(1939) 1955	300 ?	120 ?	1,500,000 ?
TOTAL		500 ?	220 ?	3,000,000 ?

APPENDIX II

SOME LINGUISTIC OBSTACLES TO MACHINE TRANSLATION*)

Y. BAR-HILLEL

Hebrew University, Jerusalem, Israel

For certain pairs of languages it has experimentally been shown that word-for-word machine translation leads to an output which can often be transformed by an expert post-editor into a quite satisfactory translation of the source text. However, if one is interested in reducing the burden of the post-editor, or if one has to do with pairs of languages for which word-for-word translation is not by itself a satisfactory basis for post-editing, it is natural to think of mechanizing the determination of the syntactic structure of the source sentences. It is a priori clear, and has again been experimentally verified, that knowledge of the syntactic structure of the sentences to be translated does considerably simplify the task of the post-editor. It is obvious, for instance, that this knowledge tends to reduce, and in the limit to eliminate, those syntactical ambiguities which are created by the word-for-word translation and which are non-existent for the human translator who treats the sentences as wholes. The task of the post-editor would then consist solely in eliminating the semantical ambiguities and in polishing up the style of the machine output. Whether these steps, too, can be taken over by machines of today or of the foreseeable future is still controversial; I myself believe that I have strong reasons for regarding it as hopeless, in general, but this is not the point I would like to discuss here.

A few years ago, I proposed what I called a <u>quasi-arithmetical notation for syntactic description</u>[1] whose employment should allow, after some refinements, for a mechanical determination of the constituent structure of any given sentence. At that time, I actually demonstrated the effectiveness of the method for relatively simple sentences only but cherished the hope that it might also work for more complex sentences, perhaps for all kinds of sentences. I am now quite convinced that this hope will not come true. As a consequence, the road to machine translation can be shown to contain more obstacles than was realized a few years ago. I think that this should be of sufficient interest to warrant some more detailed exhibition, especially since this insight is do to an important new, not to say revolutionary, view of the structure of language, recently outlined by the American linguist and logician Noam Chomsky[2], and should, in its turn and in due time, be turned into a new method of machine translation, which would be more complex than the known ones but also more effective.

Since I cannot assume acquaintance with the paper in which I introduced the quasi-arithmetical syntactical notation mentioned above, let me present it here again very briefly, with some slight modifications[3]; for a full presentation, the paper should be consulted.

- - - - - - - - - - -

*) A revised version of a talk given before the Second International Congress of Cybernetics, Namur, September 1958. It is to be printed in the Proceedings of this congress.

The basic assumption is that all words of a given language belong to one or more of the members of an infinite hierarchy of *syntactic categories*. Among these categories two are regarded as *fundamental*, viz. the categories of *nominals* and of *sentences*, denoted by *n* and *s*, respectively; the remainder are *operator categories* whose members, the *operators*, are considered as forming out of their *arguments*, always occurring to their immediate left or immediate right, more complex expressions. To illustrate: In the English sentence

John slept,

John is a *nominal*[4] and *slept* an *intransitive verbal*, i.e. an operator which out of a nominal to its left forms a sentence. We shall therefore denote the category of this operator by

n\s

(read: *n* sub *s*). In the sentence

Little John slept,

John and *slept* would belong to the mentioned categories, whereas *little* would be *adjectival*, i.e. an operator that out of a nominal to its right forms again a nominal, hence be assigned to category

n/n

(read: *n* super *n*). In

Little John slept soundly,

soundly would be an ~~(intransitive verbal)~~ *adverbial*, i.e. an operator that out of a left operator that out of a left nominal forms a sentence forms an operator that out of a left nominal forms a sentence, hence be assigned to category

(n\s)\(n\s),

or rather, to use a self-explanatory additional notational convention, to

n\s\\n\s.

Most English words, perhaps all, would belong, of course, to more than one syntactical category. *Soundly*, for instance, would belong also to n\s//n\s, to ((n\s)/n)/((n\s)/n) (think of *Belgium soundly defeated the Netherlands*), etc.

Assuming, then, that a category "dictionary" listing for each English word all its categories stands at our disposal, the task of finding out whether a given word sequence is a sentence or, more generally, a *well-formed* (or *connex*) expression and, if so, what its constituent structure is, could now be solved according to the following utterly mechanical procedure: we would write under each word of the given word sequence the symbols for all the categories to which it belongs and then start "cancelling" in all possible ways, according to either of the two following rules:

$\alpha\ \alpha\backslash\beta \rightarrow \beta$ and $\alpha/\beta\ \beta \rightarrow \alpha$.[5]

A series of such symbol sequences where each sequence results from its predecessor by one application of a cancellation rule is called a *derivation*. The last line of a derivation is its *exponent*. When the exponent consists of a single, simple or complex, symbol, the word sequence with this exponent, and with the constituent structure given by the derivation, is well-formed; if the exponent is, more specifically, *s*, the sequence is a sentence.

To illustrate, let us start with the last analyzed expression:

Little John slept soundly.

Let us assume (contrary to fact) that consultation of the category dictionary would have resulted in the following category symbol sequence:

(1) n/n   n   n\s   n\s\\n\s.

It is easy to see that there are exactly three different ways of performing the first cancellation, starting off three different derivations, viz.:

(2)   n   n\s   n\s\\n\s,

(2') n/n   s   n\s\\n\s,

(2'') n/n   n   n\s.

(2') leads into a blind alley. The other two lines, (2) and (2''), allow each for two continuations, of which one again leads into a blind alley, whereas the other allows for just one more derivation, with both exponents being s. Let me write down one of these derivations:

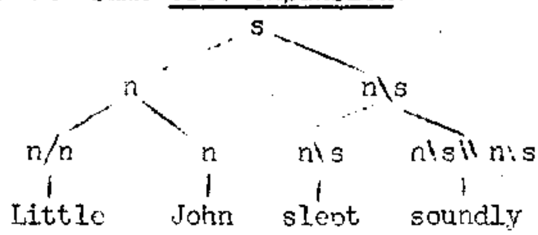
(1) n/n   n   n\s   n\s\\n\s

(2)   n   n\s   n\s\\n\s

(3)   n   n\s

(4)   s

The other derivation differs from the one just presented only in that the two cancellation steps in (2) and (3) occur in the opposite order. These two derivations are therefore equivalent, in an important sense; if fact, they correspond both to the same tree expansion:



Our second and final example will be:

Paul thought that John slept soundly.

(I hope that the somewhat shaky English of this example will be foregiven; it simplifies making the point without falsifying it.) Copying only the first entry under each word in our fictitious category dictionary, we arrive at

Paul thought that John slept soundly

n   (n\s)/n   n/s   n   n\s   n\s\\n\s.

There are two non-equivalent derivations with a single exponent. I shall again write down only one of these derivations:

(11) n   (n\s)/n   n/s   n   n\s   n\s\\n\s

(12) n   (n\s)/n   n/s   n   n\s

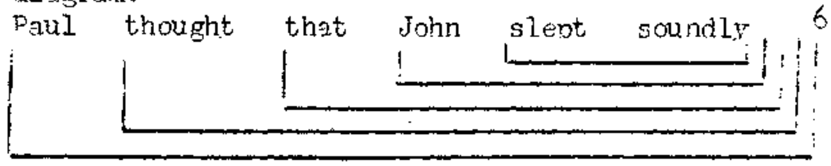
(13) n   (n\s)/n   n/s   s

(14) n   (n\s)/n   n

(15) n   n\s

(16) s

The constituent structure corresponding to this derivation can be pictured in the following parsing diagram:



6)

As said before, the situation actually is more complicated. An adequate category dictionary would contain in general more than one entry per word. *That* e.g., is often a nominal, *n*, and even more often an adjectival, *n/n*, *soundly* could as well be an *n\s//n\s* or an *((n\s)/n)/((n\s)/n)* (as mentioned above) and *thought*, finally, belongs also to categories *n*, *n\s*, *(n\s)/s* (*Paul thought John was asleep*) and, *qua* participle, to still others. It can nevertheless readily be seen that our method is capable, at least in certain cases, to determine by purely mechanical operations the specific category to which a given word belongs in its given linguistic context. In our example, e.g., listing all the mentioned categories in column form yields the following scheme:

Paul	thought	that	John	slept	soundly
n	(n\s)/n	n/s	n	n\s	n\s\\n\s
	n	n			n\s//n\s
	n\s	n/n			((n\s)/n)/((n\s)/n)
	(n\s)/s				

It would be a tedious but wholly routine exercise to determine that out of the very many derivations corresponding to this word sequence -- notice that there are 36 initial lines alone! -- there exist only three essentially different ones with a single exponent, namely, in addition to the two above-mentioned derivations just

	Paul	thought	that	John	slept	soundly
(21)	n	(n\s)/s	n/n	n	n\s	n\s\\n\s
(22)	n	(n\s)/s	n/n	n	n\s	
(23)	n	(n\s)/s	n		n\s	
(24)	n	(n\s)/s		s		
(25)	n		n\s			
(26)		s				

I still remember my surprise a few years ago when I discovered that this constituent structure is doubtless grammatical, however wildly implausible the conditions under which it would be uttered.

So far, so good, then, But, unfortunately, the actual situation is still much more complicated. It will be necessary to distinguish various kinds of nominals, for instance, singular and plural, animate and inanimate. Some additional notational means will have to be found from which it will follow that *John slept*, *The boy slept*, *Boys slept*, *The boys slept* are well-formed but that *Boy slept*, *The John slept* are not, that *The little boy slept* is connex but not *Little the boy slept*. These, and thousands other *additional refinements*,[7] can probably still be introduced without blowing the whole method up. But there are many features which make it highly doubtful whether English grammar -- or that of any other natural language, for that matter -- can at all be forced into the straitjacket of the *immediate-constituent model* and remain workable and revealing. Since the arguments against such a possibility have already been presented elsewhere with great force,[8] I shall not repeat them here in all their generality but restrict myself to the point of view of machine translation.

It takes but little to realize that the four categories mentioned above for *thought* are far from being exhaustive. In addition to its being a participle, which has already been mentioned, there are such phrases as *thought processes*, *thought thirsty* (not common but definitely grammatical), *thought provoking*, etc. In order to take care of the first two contexts, e.g., we would have to assign *thought* also to the categories *n/n* and *n/n//n/n*. ("In these contexts, *thought* occurs in the function of an adjective or an adverb, respectively" would have been one traditional way of putting the issue.) The third context would have raised the notoriously difficult problem of the status of the participle present, in addition. The task of preparing a category list that would work for all these and innumerably many other contexts is certainly much harder than the first successful analyses caused us to believe. Would not the required list become so long that the mechanical determination of the constituent structure of say, a 30-word sentence with three or four categories per word, on the average, might well require trillions of machine operations, hence be totally impractical for machines of today as well as of tomorrow?

And what with a sentence such as *Playing cards is fun*? On first sight, it seems that one has to arrive at the category *n* for the phrase *playing cards*. However, it is intuitively clear that this should not be derived from *cards* being an *n* and *playing* being an *n/n* (and not only intuitively so: notice that the next word is *is* and not *are*; *playing cards* is in our context a singular nominal). There are, of course, many other ways of enforcing an assignment of *n* to *playing cards*, but none of these, to my knowledge, is such that it would not introduce unwarranted and counter-intuitive syntactical resolutions of other sentences. "Hocus-pocus" linguistics -- as certain linguistic methods were called whose only purpose was to save certain phenomena, without regard to any intuitive (or psychological) realities -- would in our case definitely refute itself by saving also phenomena that are non-existing.

And what about a sentence like *He gave it up*? What category would *up* have to be assigned to in order that this sequence should turn out to be connex? We all feel that *gave* and *up* somehow belong together and that this is so without regard to the length of the expression that separates them. This, however, is definitely beyond the reach of the immediate constituent model in which the immediate constituents of a connex expression are always contiguous or, to put it in a different terminology, where modified expression and modifying expression have to stand one directly after, or before, the other.[9)]

If now the immediate constituent model is not good enough to serve as a general model for the whole grammar of a given language, the method of mechanical structure determination outlined above can no longer be assumed to be of general validity, either. As a matter of fact, I had noticed already six years ago that complex sentences could not be analyzed well by this method as it stood then but I had rather hoped that this was due only to lack of refinement. I have now come to realize that its failure in the more complex cases has a much deeper cause: the linguistic model on which this method was based is just not good enough.

Since the thinking of the linguists working on machine translation was mostly governed by the immediate constituent model, unless they were working with a still more primitive model, a communication-theoretical finite-state Markov process model (or, of course, working without any model), it should not be really surprising that so little progress was made during the last years in the mechanization of the syntactic analysis of languages. I, for one, am satisfied with this explanation of the present stagnation in this respect.

Having identified the nature of this obstacle to machine translation, we must, of course, ask ourselves what consequences are to be drawn from this identification for future work on MT. The answer is rather simple as such, though its exact implications are far from being clear. A better model for the working of grammar, i.e. for the synthesis of well-formed expressions, especially sentences, out of the linguistic elements -- which, for MT purposes, are the letters and other elementary graphic signs such as numerals, punctuation marks, etc. -- has first to be set up and then turned around to allow for the mechanical analysis of the resulting large units. Chomsky and Harris[10] have shown us outlines of a third, more powerful model for linguistic synthesis, the so-called transformational model. It does not discard the immediate-constituent model but rather supplements it. The former model remains intact for a certain kind of simple sentences, the so-called kernel sentences (or rather for their underlying terminal strings)[11] -- and our method of mechanical structure determination remains therefore valid for these sentences --, but has to be supplemented by additional procedures, the so-called transformations, in order to account for the synthesis of all sentences.

The answer to the question, "What is the constituent structure of the sentence, He gave it up?", is now: this sentence has no proper constituent structure; it is the result of a certain transformation on the terminal string, He gave up it, which has indeed a rather simple and perspicuous constituent structure. The answer to the question, "What is the subject of the sentence, Playing cards is fun?", is now -- whatever grammarians had to say on this topic until now (and what they had to say was highly unsatisfactory and often contradictory) -- that this sentence, not being a terminal string, has no proper subject but is rather the result of certain transformations on certain terminal strings.[12] (The actual situation is too complicated to be treated in the space at my disposal.)

Each sentence, according to our last model, is then the result of a series of one or more transformations performed one after the other on one or more terminal strings -- unless, of course, it is a terminal string itself. A complete analysis, mechanical or otherwise, of a given sentence has to tell us what its basic terminal strings are, together with their constituent structure, and what transformations, and in what order, were performed upon them. Assuming that a complete transformational grammar, for some given language, has been prepared, the preparation of a corresponding analytical (or operational) grammar is a formidable, though perhaps not necessarily an impossible task. So far, of course, no transformational grammar exists for any language, to any serious degree of completeness.

The recognition that immediate constituent grammars have to be supplemented by transformational grammars makes the task of mechanizing translation look much harder, but the resulting picture is not at all uniformly black. On the contrary, there are reasons to suppose that the additional insight we get on the basis of this model will not only be of decisive importance for theoretical linguistics but may well turn out to facilitate the mechanization of translation from new angles.

First: you remember that one of our previously analyzed sentences was

Paul thought that John slept soundly

and the troubles we foresaw in its mechanical analysis. It is obvious, however, that in a transformational grammar this sentence will not be a terminal string but rather (1) either the result of a certain kind of "fusing" transformation on the sequence of the two terminal strings

Paul thought this: John slept soundly

or (2) the result of two transformations, the first being the same "fusing" transformation performed, however, on somewhat different terminal strings

Paul thought this: That John slept soundly,

the result of which would be

Paul thought that that John slept soundly,

the second transformation being a certain kind of "elliptic" transformation causing, in our case, the omission of the first <u>that</u>.

No longer, then, will $(\underline{n}\backslash\underline{s})/\underline{s}$ be regarded as one of the categories of <u>thought</u>, nor $\underline{n}/\underline{n}$ and $\underline{n}/\underline{n}//\underline{n}/\underline{n}$, as <u>thought processes</u> and <u>thought thirsty</u> will now be treated as resulting from <u>processes of thought</u> and <u>thirsty for thought</u> by certain transformations.

The first gain consists, then, in that the number of categories per word will almost always be less, sometimes much less, than under the former method. For some words this number will now be zero, indicating that no sentence containing such words is a terminal string. To give an example: <u>sleeping</u> will not be assigned to any category, any sentence containing this word being considered as the result of a transformation. (<u>Interesting</u>, however, will be assigned to the category $\underline{n}/\underline{n}$.)[14] That there might be words which do not belong to any syntactic category will strike many linguists as rather queer, but I am convinced that on second sight they will realize the enormous advantages of such an attitude; innumerable pseudo-problems have in the past been created by the search for the syntactic category (the traditional term is, of course, "part of speech") of certain words or phrases which -- under the new model -- just don't belong to any category. This is -- if I may be allowed one generalization -- just one more instance of the very common class of situations where the attempt of applying a model which is very useful within certain limits leads, when pushed beyond these limits, to pseudo-problems and their pseudo-solutions.

The second gain is somewhat more speculative: it seems likely, but has so far not been seriously tested, that languages will be much more similar with regard to their terminal string structure than with regard to the structure of the totality of their sentences. Word-for-word translation of terminal strings, with some occasional permuting, seems to yield satisfactory results for many pairs of languages, including those for which this kind of translation does not work at all with regard to more complex sentences.

The most remarkable gain, however, would be achieved when it turned out that between the sets of transformation of two languages there existed a close semantic relationship. Should it happen that for certain two languages, L_1 and L_2, there exist two transformations, say t_1 and t_2, such that for any semantically equivalent terminal strings of these languages, k_1 and k_2, $t_1(k_1)$ is semantically equivalent to $t_2(k_2)$, this would allow for a relatively simple mechanization of the translation, provided, of course, that the syntactic analysis of L_1 has been mechanized, whereas a word-for-word translation of $t_1(k_1)$ into L_2 might be highly unsatisfactory.

Of course, there is but little hope that the sets of transformations of two languages which do not stand in any close genetical relationship will do us the favor of exhibiting isomorphism or near-isomorphism with regard to semantic equivalence. So far, there exists to my knowledge no general theory of machine translation which would ensure that, if only the precepts of this theory are followed, the target language counterpart (or counterparts) of any sentence of a given source language will be no more and no less syntactically ambiguous than the original sentence itself. Current statements to the contrary seem to me palpably false, and any hope for an imminent establishment of such a theory -- unsubstantiated. Great progress has been made in this respect with regard to certain ordered pairs of languages, such as French-English, German-English, Russian-English, English-Russian, German-Russian and French-Russian, partly prior to the appearance of the transformational model and without any conscious use of its methods, and more progress may be expected in the future through a conscious use of these methods. As one necessary condition for further success I regard the recognition on behalf of the workers on MT that the model with which they were working, consciously or unconsciously, during the first decade of their endeavors was too crude and has to be replaced by a much more complex but also much better fitting model of linguistic structure.

NOTES

1) "A quasi-arithmetical notation for syntactic description", Language 29:47-58 (1953).

2) Noam Chomsky, Syntactic Structures, 's-Gravenhage, 1957.

3) These modifications refer both to terminology and to notation. The latter are influenced by J. Lambek, "The mathematics of sentence structure", American Mathematical Monthly 65:154 (1958).

4) Nominals, verbals, adjectivals, etc., in my usage, are syntactical categories. They should not be confused with nouns, verbs, adjectives, etc., which are morphological (paradigmatic) categories in my usage. The connection between these two classifications, as the choice of terms is intended to indicate, is that nouns usually, though by no means always, belong to the syntactical category of nominals, etc., and that most expressions belonging to the syntactical category of nominals of course only if they are single words, are nouns.

5) The reading of these rules should be self-explanatory. The first, for instance, reads: Replace the sequence of two category symbols, the first of which is any category symbol whatsoever and the second of which consists of the first symbol followed by a left diagonal stroke followed by any category symbol whatsoever, by this last category symbol.

6) The other single exponent derivation yields a constituent structure whose diagram is

Paul thought that John slept soundly .

If this structure is regarded as unacceptable, the notation will have to be considerably refined in order to exclude this derivation.

With regard to the problems arising in connection with the fact that the notation (n\s)/n creates an arbitrary-looking preferential reading of what should "naturally" have been written n\s/n, see op. cit. in note 1, p. 55 and op. cit. in note 3. Both treatments do not yet cover all aspects of the problem.

7) Such as the one discussed in the preceding note.

8) In op. cit. in note 2, as well as in other recent publications by the same author.

9) Discontinuous constituents were occasionally discussed in theoretical linguistics, but not before Chomsky was it realized what a difference this makes as against continuous and contiguous constituents.

10) Zellig S. Harris, "Co-occurrence and transformation in linguistic structure", Language 33:283-340 (1957).

11) Cf. op. cit. in note 2, p. 45.

12) With appropriate safeguards, but only with such safeguards, one might also answer the first question by saying that the sentence, He gave it up, has He and

gave it up as its immediate constituents, and that its second component has the discontinuous expression gave...up and it as its immediate constituents. Likewise, the answer to the second question could also be formulated by saying that Playing cards is its quasi-subject, but this requires, of course, a prior definition of 'quasi-subject'.

13) This is only a first approximation. Actually, a satisfactory description will have to be much more complex.

14) Why? Hint: we have very interesting but not very sleeping.

APPENDIX III

DECISION PROCEDURES FOR STRUCTURE IN NATURAL LANGUAGES*)

Y. BAR-HILLEL

Hebrew University, Jerusalem, Israel

The rules of formation of a logistic system are by definition[1] such that the notion of formula, well-formed formula or sentence, determined by these rules, is effectively decidable. However, I am not convinced that the arguments brought forth by Church[2] to the effect that sentencehood has to be an effectively decidable notion for any system that may be used for communication purposes are conclusive. I therefore regard it to be a serious problem whether the syntactic structure of a natural language such as English can always be adequately described by a set of formation rules that guarantee the decidability of the notion of sentence or, for that matter, of any other syntactical structures such as phrases etc. Inasmuch as there exist good reasons for doubting whether the answer to this problem is affirmative, the prospects for fully-automatic, high-quality translation from one natural language into another natural language look dimmer than many workers in the field of machine translation would like to think. This is so since not even one necessary, though by no means sufficient, condition for this process, namely the mechanical determination of the syntactical structure of any given sentence in the source language, could possibly be completely fulfilled. Though applicability to machine translation is often in the back of my thinking on the description of the syntax of natural languages, I shall refer here no longer to this application, having dealt with it elsewhere at some length.[3]

The seriousness of our problem has apparently not been sufficiently recognized so far because many linguists explicitly, and most if not all of them as well as most logicians implicitly, believed that the syntactical structure of natural languages is adequately describable by an immediate constituent model, or a phrase structure model according to the term recently introduced by Chomsky.[4] It is indeed true that if natural languages were adequately describable in terms of such a model, there would exist a decision procedure for structure, as I have shown in effect, though not with full rigor, in a paper published six years ago.[5]

Before I proceed to present some arguments for the fact that the phrase structure model is not fully adequate, let me spend some time in presenting again,

- - - - - - - - - -

*) A revised version of a talk given before the Colloque de Logique, Louvain, September, 1958. The present version was published in the Belgian journal Logique et Analyse, N.S., 2^e Année, No. 5, Janvier 1959. Since, however, this issue was sent to the printers only in the second week of February 1959, according to a communication from its editors, I decided not to wait for the arrival of the reprints and to reproduce it myself in the present form. So some minor discrepancies between the versions may be expected.

The reader will realize that the present paper overlaps with the one reproduced in Appendix II. After some hesitation, I decided nevertheless to include it here, as it is more elaborate in many points. A consolidation of my views on the theoretical aspects of MT is in preparation.

in briefer and, I hope, improved form, an informal outline of this proof. The basic idea behind the immediate constituent model is that every sentence can be regarded as a result of the operation of one continuous part of it upon the remainder such that those constituent parts which in general are not sentences themselves, but rather phrases, are themselves again the product of the operation of some continuous part upon the remainder, etc., until one arrives at the final constituents, say words or morphemes. To illustrate:

Young John slept soundly

would be regarded as the result of the operation of slept soundly upon young John; slept soundly in its turn would be considered the result of the operation of soundly upon slept and young John the result of the operation of young upon John. All this so far is nothing but reformulation in somewhat unfamiliar terms of the procedure well known from school days as parsing. As linguists put it, young John and slept soundly are the immediate constituents of the sentence under discussion, young and John the immediate constituents of the first immediate constituent of the sentence, slept and soundly the immediate constituents of the second immediate constituent. Hence altogether young, John, slept and soundly are the final constituents of the given sentence.

Another basic feature of the model is that all operator constituents must be contiguous with their argument constituents. Both these features are exemplified in our illustration, but this of course is by no means a proof that this model can be carried through all of language. On the contrary, linguists have realized that occasionally discontinuous constituents have to be taken into account, but they seem to have believed that these were exceptions which did not seriously affect the validity of the model with which they were used to work.

In most language systems invented by logicians, the two mentioned features were automatically incorporated into their respective rules of formation. The problems arising in connection with discontinuous expressions were, to my knowledge, never explicitly discussed by logicians.

According to the immediate constituent model, every word -- and we shall for our purposes consider words to be the basic syntactical elements -- of a natural language belongs to one or more syntactical category. Among these categories some will be pure argument categories, by which term I denote a category whose members always serve as arguments and never as operators, as well as operator categories whose members may operate upon other words though they may perhaps also be operated upon by other operator expressions. John, for instance, inasmuch as it belongs to the syntactic category of nominals, is always an argument and never an operator. Slept, inasmuch as it belongs to the category of intransitive verbals, may operate upon a nominal such as John to form the sentence John slept, but may also be operated upon by the adverbial soundly to form the intransitive verbal expression slept soundly. A word may belong to more than one category not only because it may be regarded as homonymous -- as would be the case with regard to sleep, which clearly belongs to the category of nominals as well as to the category of intransitive verbals -- but also because, for instance, many adverbials operate upon intransitive verbals as well as upon transitive verbals: soundly, for example in the sentence

Belgium soundly defeated the Netherlands

(in the last soccer game, of course), operates upon the transitive verbal defeated, forming the transitive verbal expression soundly defeated, and has therefore a different kind of argument as well as a different kind of value than has soundly when operating upon slept.

In order to exhibit the decision procedure for constituent structure let us denote, following Leśniewski and Ajdukiewicz, the category of nominals by '<u>n</u>' and the category of declarative sentences by '<u>s</u>'. (Since I am engaged in presenting an outline only, I shall not go here into the very difficult question to what degree these two argument categories would have to be refined and expanded in order to get even the beginnings of a reasonably working model.) Operator categories will be denoted by symbols that will indicate both the categories of their arguments and the category of the resulting expression. In addition, since arguments may be positioned either at the immediate left or at the immediate right of their operator, these positions too will have to be indicated in the symbolism. Therefore, I shall, for instance, denote the category of <u>slept</u> by '<u>n\s</u>' -- read: <u>n</u> sub <u>s</u> -- and the category of <u>young</u> by '<u>n/n</u>' -- read: <u>n</u> super <u>n</u>,[6] -- where the direction of the slash indicates in an obvious fashion whether the argument is to the left or to the right. <u>And</u>, for instance, <u>qua</u> sentence connective, will be assigned to the category <u>s\s/s</u>*) since in this function it is a word that out of a sentence to its immediate left and a sentence to its immediate right forms a sentence. <u>Soundly</u> will belong to the categories (<u>n\s</u>)\(<u>n\s</u>) -- to be abbreviated in a self-explanatory way as <u>n\s\\n\s</u> -- and <u>n\s/n//n\s/n</u> -- as well as well as to a few other categories.

Assume now that we have a complete category list of all English words, i.e. a list which gives all the syntactical categories to which every English word may belong. In order to arrive by a completely mechanical procedure at the constituent structure of any given English sentence, one would only have to copy from the category list the category symbols for all the words in this sentence, write them down in columns and go to work on them according to the following rule: <u>Replace a sequence of three symbols, having respectively the form α, $\alpha\backslash\beta/\gamma$ and γ</u> with β. This rule comprises as limiting cases the following two subrules:

(1) Replace the sequence of symbols of the form α and $\alpha\backslash\beta$ by β.
(2) Replace the sequence of symbols of the form β/γ and γ by β.

Instead of going into a detailed but rather obvious description of the decision procedure let us illustrate through a somewhat more elaborate example. Assume that the word sequence to be tested for sentencehood as well as for its constituent structure is

<u>Paul thought</u> that <u>John slept soundly</u>.

Assume further that copying from the categorylist yields the following result:

<u>Paul</u>	<u>thought</u>	<u>that</u>	<u>John</u>	<u>slept</u>	<u>soundly</u>
n	n	n	n	n\s	n\s\\n\s
	n\s	n\n			n\s/n//n\s/n
	n\s/n	n\s			⋮
	n\s/s				
	⋮				

(the three dots indicating that the complete list would probably contain further entries which shall, however, be here disregarded for the sake of simplification). The reader will do well to envisage contexts in which <u>thought</u> and <u>that</u> will belong to each of the given categories. He might as well try to find out to

*) [Added for the present version:] This notation may turn out to be too lax for certain purposes. A more strict notation is (s\s)/s. Similarly, the main rule given in the following paragraph should officially always be replaced by the two subrules given there. The explanation of a derivation, given below, is therefore somewhat inaccurate, and so are the examples. There should be no difficulty in introducing additional rigor, when required, in accordance with the procedure followed in Appendix II.

which categories thought would belong in such contexts as: John had thought of..., ...thought processes, and ...thought provoking... .

Now taking into account only the categories explicitly indicated we have twenty-four initial symbol sequences to which we will apply our rule. Starting for instance with

```
n   n   n   n   n\s   n\s\\n\s
```

we see that subrule (1) can be applied for the fourth and fifth symbols yielding s. The resulting sequence is now

```
n   n   n   s   n\s\\n\s,
```

which obviously cannot be further operated upon. The same subrule operating upon the fifth and sixth symbols yields n\s, hence the sequence

```
n   n   n   n   n\s,
```

which has once more to be operated upon by the same subrule yielding

```
n   n   n   s,
```

which cannot be processed any further.

Performing these operations upon all the twenty-four initial symbol sequences through all possible continuations, we would find that there exist exactly three derivations -- as we shall call columns of symbol sequences each of which (with the exception of the first, of course) results from the preceding line by one application of the rule -- whose final line, or exponent, consists of a single symbol which in both cases is 's'.

Here are the derivations:

```
n   n\s/n   n/s   n   n\s   n\s\\n\s
n   n\s/n   n/s   n      n\s
n   n\s/n   n/s      s
n   n\s/n       n
        s
```

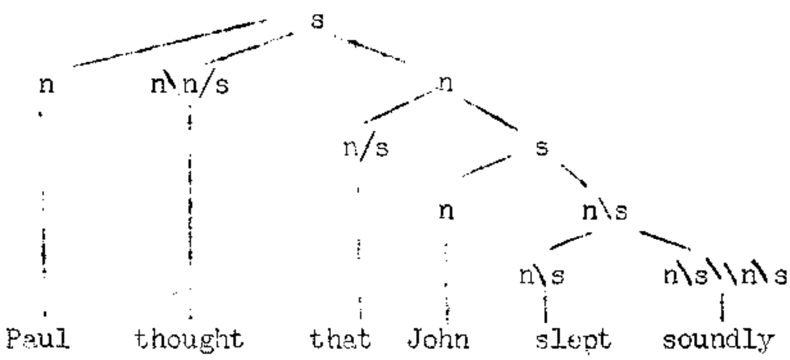
```
n   n\s/s   n/n   n   n\s   n\s\\n\s
n   n\s/s   n/n   n      n\s
n   n\s/s      n         n\s
n   n\s/s          s
        s
```

```
n   n\s/s   n/n   n   n\s   n\s\\n\s
n   n\s/s      n      n\s   n\s\\n\s
n   n\s/s      n         n/s
n   n\s/s          s
        s
```

The last two derivations being equivalent, in a rather obvious sense of the word, we have only two essentially different derivations before us, indicating, probably to the surprise of many readers -- and to my own surprise some six years ago when I came across this situation simulating a machine processing of this illustration --, that the sentence under discussion is syntactically ambiguous or constructionally homonymous. The reader will do well to read out aloud this sentence according to its two essentially different constituent structures which in this case make the sentence also semantically ambiguous as such, though one constituent structure is much less likely to be used than the other.

I hope that this illustration is sufficient to show that under the essential and, as we shall see, highly problematic assumption that a complete and completely adequate category list is available, there exists indeed a wholly mechanical procedure to determine whether a given word sequence is a declarative sentence under one of its constituent structures as well as what all of its constituent structures are.

For certain purposes it is worthwhile to look upon our derivation procedure upside down, i.e. to deal with expansion rather than with derivation. The expansion corresponding to the first derivation exhibited above of our sample sentence would look like the following tree:



(Two derivations, by the way, are equivalent if they correspond to the same tree.)

How well then does the immediate constituent model work? Apparently quite well for relatively short sentences such as those discussed so far, but even there not too well. The number of categories to which the English words will have to be assigned to make the category list reasonably adequate will occasionally have to be rather large, and the categories themselves rather complex. In addition, it is quite clear that not only will one have to work with highly complex refinements of the categories mentioned so far in order to take care, for example, of the fact that John sleeps is a sentence but not John sleep, but that one will also hav to refine the category of sentences and distinguish between declarative sentences, imperative sentences, yes-or-no question sentences, wh-question sentences, etc., these various types not being reducible to each other under our model. These refinements may result in such a piling up of category symbols assigned to the words occurring in a given sentence that the number of derivations would easily run into the trillions, hence be beyond the practical capacity of even the fastest electronic computers. For instance, if the average number of categories of the twenty words of a given English sentence is four, we will have up to 4^{20} initial lines and a still enormously higher number of derivations. This means, then, that the indicated method of mechanically resolving the syntactical structure of any given English sentence would certainly be impractical as such. However, were it the case that this is still a theoretically adequate method, one could think of certain improvements which would reduce the required number of operations by many orders of magnitude. Unfortunately, however, the actual situation seems to be much worse. It is not only a matter of practicality, but it seems that the whole model is just not good enough. Already six years ago I was worried by sentences such as

John, unfortunately, slept soundly

which, so it appears at least, cannot be handled by a model incorporating the two above-mentioned basic features. Notice that there is no trouble with the slightly different and semantically, though perhaps not stylistically, equivalent sentence

Unfortunately, John slept soundly.

Assigning unfortunately to the category s/s, a wholly natural and intuitive assignment, we arrive at an adequate syntactical analysis. This assignment, however, clearly does not work for John, unfortunately, slept soundly, as the reader will easily verify for himself. It is of course possible that some other less natural category assignment to unfortunately, perhaps combined with some ingenious treatment of the commas (which so far have been completely disregarded in the immediate constituent model), would do the trick. It seems, however, unlikely that such an assignment could be made in a fashion which would not be almost entirely ad hoc. And this would not only be esthetically and methodologically repugnant but also, in all likelihood, have unpleasant repercussions inasmuch as word sequences which intuitively would not be regarded as grammatical sentences would have derivations with an exponent of s.

A similar situation, but even simpler since no commas are involved, arises with regard to the word sequence

He looked it up.

Regarding he and it as belonging to the categories n -- leaving aside once more the clearly required refinements --, looked as belonging to the category n\s/n, as seems natural, it seems highly implausible that any category assignment of up which would not be woefully ad hoc would insure the sentencehood of the given word sequence. Assigning up, for instance, to the category s\s would obviously result in a derivation with an exponent s, but this unnatural saving of the phenomena would immediately retaliate with the unwanted imposition of sentencehood to such sequences as

He went home up.

(For further examples of the breakdown of the phrase structure model see Chomsky's Syntactic Structures,[7] to which I owe much of the present argument.)

Every English speaker, I presume, feels that in our sentence

He looked it up

looked and up belong somehow together. Indeed there is no trouble with such a sequence as

He looked up this argument,

as the reader will easily verify for himself, if only up is assigned in a completely intuitive fashion to the category n\s/n\\n\s/n. This being so, assigning up to a different category, whatever it now may be, in the sentence

He looked it up

looks now even more artificial than before.

These simple facts indicate, though it cannot be said that they prove in the strong sense used in mathematics, that the immediate constituent model is not an adequate one as such, but has to be supplemented in one way or another.

Let me finish this discussion by presenting a very brief outline of one such supplementation method, referring the reader for a fuller discussion to Chomsky's mentioned book and other publications of his[8]. The new model, called the transformational model, assumes that sentences are generated not only by the procedure we called above expansion, but also in addition by so-called transformations. One such transformation, for instance, would transform the so-called terminal string of the following expansion

```
                          s
          ________________|________________
         |                |                |
         n             n\s/n               n
         |            /      \             |
         |       n\s/n    n\s/n\\n\s/n     |
         |         |           |           |
        He       gave         up          it
```

i.e. He gave up it, which is, of course, not an English sentence, into He gave it up by a certain obligatory transformation. This transformation rule which states in effect that in certain environments certain word sequences have to be turned around is clearly beyond the reach of an immediate constituent model. On the other hand, this way of looking at how the sentence He gave it up was generated has a rather natural appearance, and might well correspond, at least in spirit, to the way old-fashioned, traditional grammar has dealt with the situation.

Other transformations transform two terminal strings into one sentence. One of these, an optional one, would operate upon the sequence of the two terminal strings (which are in this special case sentences in their own right)

Paul thought it. John slept soundly.

and turn this sequence into the sentence

Paul thought that John slept soundly.

This very same transformation operates upon the sequence

Paul thought it. That John slept soundly.

and transforms it into

Paul thought that that John slept soundly.

Yet another transformation to the effect that under certain determined conditions that may be omitted would transform this last sentence into

Paul thought that John slept soundly.

This way of looking at the situation results now in a natural and adequate explanation of the constructional homonymy of the last sentence. We also realize, by the way, that transformations may operate upon the results of prior transformations.

Linguists, such as Harris, Chomsky, and their associates, who are at work at the development of this new kind of model[9] have already unveiled a large number of transformations amounting to many hundreds in English. It is, however, quite clear that the transformations introduced so far are not yet sufficient to account for all intuitively possible English sentences. It is at this state that the question mentioned at the beginning of this paper arises -- whether there exists a decision procedure for structure in English, or in other natural languages for that matter, since it is unlikely that the natural languages should differ among themselves in this respect. Obviously the answer to our question will depend upon the exact nature of the transformations. Only when we will have a better and more extensive understanding of the kind of transformations at work, will we be in a position to fruitfully attack our problem. At this moment one could only speculate about this answer, and it is doubtful whether such speculations would be worthwhile. In any case, even the possiblity that for a certain set of formation rules in English the notion of English sentence would not be a decidable (or general recursive) one seems exciting enough to warrant an increase in interest in our problem among mathematical logicians who by training are in many respects in a better position to attack it than are linguists. Chomsky has already been able to show that there exist highly interesting connections between the theory of linguistic models and such theories as the theory of automata, recursive function theory (perhaps especially conspicuous in the form of the theory of algorithms) and the theory of Post canonical systems. This multiple relationship indicates that we have in all probability in the theory of language models an interesting new field in which cross-fertilization of mathematical logic and structural linguistics should lead to important results.

NOTES

1) See, e.g., A. Church, *Introduction to mathematical logic*, I, Princeton, 1956, p. 51. There exist, however, less demanding conceptions.

2) *Ibid.*, p. 53.

3) In "Some linguistic obstacles to machine translation", forthcoming in the *Proceedings of the Second International Congress of Cybernetics*, held in Namur September 1958.

4) See N. Chomsky, "Three models for the description of language", *IRE Transactions on Information Theory*, Vol. IT-2, No. 3 (1956) and *Syntactic structures*, 's-Gravenhage, 1957.

5) "A quasi-arithmetical notation for syntactic description", *Language* 29:47-58 (1953).

6) See K. Ajdukiewicz, "Die syntaktische Konnexitaet", *Studia Philosophica* 1:1-27 (1935-36); cf. A.A. Fraenkel and Y. Bar-Hillel, *Foundations of set theory*, Amsterdam, 1958, pp. 169-170.

7) In the paper mentioned in note 5, I used a less convenient symbolism. The present symbolism is due to J. Lambek, "The mathematics of sentence structure", *American Mathematical Monthly* 65:154 (1958).

8) See above, note 4.

9) Viz., to those mentioned above in note 4, as well as, for instance, to a forthcoming paper, "A transformational approach to syntax".

10) In addition to Chomsky's publications, see Z.S. Harris, "Cooccurrence and transformations in linguistic structure, *Language* 33:283-340 (1957) and the excellent review of Chomsky's *Syntactic structures* by R.L. Lees in *Language* 33:375-408 (1957).

APPENDIX IV

A DEMONSTRATION OF THE NON-FEASIBILITY OF FULLY-AUTOMATIC HIGH-QUALITY MACHINE TRANSLATION*

Y. BAR-HILLEL

Hebrew University, Jerusalem, Israel

One of the reasons why we do not as yet have any translation centers, not even in the planning stage, in which electronic computers, general or special purpose, are used to automate certain parts of the translation process, in spite of the fact that such centers would fulfill a vital function in saving a considerable amount of qualified human translator time per document translated, and thereby facilitate more, quicker and, after some time, cheaper translation, is the reluctance of many MT workers to recognize that their pet idea of inventing a method for fully-automatic high-quality machine translation (FAHQMT) is just a dream which will not come true in the foreseeable future. By not realizing the practical futility of this aim, whatever its motivational importance for certain types of basic research, they have managed to fool themselves and the agencies which sponsored their research not to be satisfied with a partly automated translation system whose principles are well understood today, but to wait for the real thing which was believed, and made believe, to be just around the corner.

During the last year, I have repeatedly tried, through personal talks, lectures before conferences and small groups, as well as in articles[1], to point out the illusoriness of the FAHQMT ideal already in respect to mechanical determination of the syntactical structure of a given source language sentence. These efforts of mine were based on certain deep theoretical insights into linguistic structure recently obtained by Chomsky[2]. Today, I shall show that there exist extremely simple sentences in English -- and the same holds, I am sure, for any other natural language -- which, within certain linguistic contexts, would be uniquely and unambiguously translated into, say, French or German or Russian or what have you by anyone with a sufficient knowledge of the two languages involved, though I know of no program that would enable a machine to come up with this unique rendering unless done so by a completely arbitrary and ad hoc procedure whose futility would show itself in the next example. I defy any of the MT experts gathered here -- and anybody else -- to show me where I am wrong. In case they are unable to do this, let me suggest that they stop talking about the attainability of FAHQMT, and thereby give the green light to those agencies and people who are interested in overcoming the severe problems created by the shortage of qualified human translators. If they themselves are not willing to cooperate in the establishment of a working, partly-automated, high-quality translation outfit because of the small amount of intellectual satisfaction that will accompany

- - - - - - - -

*) This paper was prepared under a more general research program supported by the Office of Naval Research, Information Systems Branch, Contract No. Nonr-2578(00), NR 049-130. It is to be read before the International Conference for Information Processing, Paris, June 1959.

such an achievement, I for one understand this attitude fully and even share it to such a degree that I have no intention to spend my time on such a project. Let them, and myself, by all means go on and investigate the countless number of basic problems with a real challenge in them, concerning language models, special-purpose program languages for translation, machines that will learn to translate, etc. But I do not think they should discourage other people eager to establish a system, any system, that does save valuable man-power and does solve an urgent problem, by threatening them, so to speak, with the outlook of having to face tomorrow an incomparably better system that will make their efforts and achievements look childish and pointless.

I now come to my sentence. It is just:

The box was in the pen.

The linguistic context from which this sentence is taken is, say, the following:

Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

Assume, for simplicity's sake, that 'pen' in English has only the following two meanings: (1) a certain writing utensil, (2) an enclosure where small children can play. I now claim that no existing or imaginable program will enable an electronic computer to determine that the word 'pen' in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this "automatically". Incidentally, we realize that the issue is not one that concerns translation proper, i.e., the transition from one language to another, but a preliminary stage of this process, i.e., the determination of the specific meaning in context of a word which, in isolation, is semantically ambiguous (relative to a given target-language, if you so wish).

It is an old prejudice, but nevertheless a prejudice, that taking into consideration a sufficiently large linguistic environment as such will suffice to reduce the semantical ambiguity of a given word. Let me quote from the memorandum which Warren Weaver sent on July 15, 1949 to some two hundred of his acquaintances and which became one of the prime movers of MT research in general and directly initiated the well-known researches of Reifler and Kaplan[3]: "...if...one can see not only the central word in question, but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning of the central word. The formal truth of this statement becomes clear when one mentions that the middle word of a whole article or a whole book is unambiguous if one has read the whole article or book, providing of course that the article or book is sufficiently well written to communicate at all." Weaver then goes on to pose the practical question: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word", a question which was, we recall, so successfully answered by Kaplan. But Weaver's seemingly lucid argument is riddled with a fateful fallacy: the argument is doubtless valid (fortified, as it is, by the escape-clause beginning with 'providing') but only for intelligent readers, for whom the article or book was written to begin with. Weaver himself thought at that time that the argument is valid also for an electronic computer, though he did not say so explicitly in the quoted passage, and on the contrary, used the word 'one'; that this is so, will be clear to anyone who reads with care the whole section headed "Meaning and Context". In this fallacious transfer

Weaver has been followed by almost every author on MT problems, including the Russian ones. It would be very easy to provide as many quotations as you wish to corroborate this statement of mine. But this is probably unnecessary since I do not believe that someone would wish to challenge me on this point.

Now, what exactly is going on here? Why is it that a machine with a memory capacity sufficient to deal with a whole paragraph at a time, and a sytactico-semantic program that goes, if necessary, beyond the boundaries of single sentences up to a whole paragraph (and, for the sake of the argument, up to a whole book, if you so wish) -- something which has so far not gotten beyond the barest and vaguest outlines -- is still powerless to determine the meaning of 'pen' in our sample sentence within the given paragraph? The explanation is extremely simple, and it is nothing short of amazing that, to my knowledge, this point has never been made before, in the context of MT, though it must surely have been made many times in other contexts. What makes an intelligent human reader grasp this meaning so unhesitatingly is, in addition to all the other features that have been discussed by MT workers -- Dostert[4], e.g., lists no less than seven of what he calls areas of meaning determination, none of which, however, takes care of our simple example --, this intelligent reader's <u>knowledge</u> that the relative sizes of pens-qua-writing-utensils, toy boxes and <u>pens-qua-play-pens</u> are such that when someone writes under ordinary circumstances and in something like the given context, "The box was in the pen" (and the occurrence of this sentence in the mentioned paragraph tends to increase the confidence of the reader that the circumstances are ordinary, though the whole paragraph could, of course, still have formed part of a larger fairy tale, or of some dream story, etc.), he almost certainly refers to a play-pen and most certainly not to a writing pen. This knowledge stands at the disposal of the average human reader beyond a certain age, and the writer takes this into account. This knowledge does not stand at the disposal of the electronic computer and none of the dictionaries or programs for the elimination of polysemy puts this knowledge at its disposal.

Whenever I offered this argument before one of my colleagues working on MT, their first reaction was: "But why not envisage a system which will put this knowledge at the disposal of the translation machine?" Understandable as this reaction is, it is very easy to show its utter futility. What such a suggestion amounts to, if taken seriously, is the requirement that a translation machine should not only be supplied with a dictionary but also with a universal encyclopedia. This is, however, surely utterly chimerical and hardly deserves any further discussion. Since, however, the idea of a machine with encyclopedic knowledge has popped up also on other occasions, probably also during the present conference, let me add a few words on this topic. The number of facts we human beings know is, in a certain very pregnant sense, infinite. Knowing for instance, that at a certain moment there are exactly eight chairs in a certain room, we also know that there are more than five chairs, less than 9, 10, 11, 12, and so on <u>ad infinitum</u>, chairs in that room. If you so wish, we know all these additional facts by inferences which we are able to perform, at least in this particular case, instantaneously, and it is clear that they are not, in any serious sense, stored in our memory. Though one could envisage that a machine would be capable of performing the same inferences, there exists so far no serious proposal for a scheme that would make a machine perform such inferences in the same or similar circumstances under which an intelligent human being would perform them. Though a lot of thought should surely be given to the problems which could only be touched here very little, it would very definitely mean putting the horse before the cart if practical MT would have to wait for their solution. These problems are clearly many degrees of order more difficult than the problem of establishing practical machine aids to translation. I believe that it is of decisive importance to get a clear view of this whole issue and hope that my remarks will contribute to its clarification.

I have no idea how often sentences of the mentioned kind, whose ambiguity is resolvable only on the basis of extra-linguistic knowledge which cannot be presumed to be at the disposal of a computer, occur on the average in the various types of documents in whose translation one might be interested. I am quite ready to assume that they would occur rather infrequently in certain scientific texts. I am ready to admit that none might occur on a whole page or even in some whole article. But so long as they will occur <u>sometimes</u>, a translation outfit that will claim that its output is of a quality comparable to that of a qualified human translator will have to use a post-editor. As soon as this is granted, the greatest obstacle to practical MT has been overcome, and the way is free for an unprejudiced discussion of the best human use of the human partner in the translation outfit.

Having shown, I hope, that FAHQMT is out of the question for the foreseeable future because of the existence of a large number of sentences, the determination of whose meaning, unambiguous for a human reader, is beyond the reach of machines[5], let me now discuss this issue of reduction of semantical ambiguity a little further. There exist in the main two methods of reducing semantical ambiguity. One is the use of idioglossaries, the other is the already mentioned method of utilizing the immediate linguistic environment of the word which is ambiguous in isolation. Though some doubts have been raised on occasion as to the validity of the first of these methods, I do not know of any serious attempt to put its validity to test. At this point I would only like to impress you with the vital necessity of performing this test before an MT method based upon its utilization is claimed to yield high-quality translations, even in collaboration with a post-editor. It is just the great effectiveness of the use of idioglossaries in general which is apt to yield disastrously wrong translations on occasion without giving the post-editor even a chance to correct these mistakes. It is just because a certain Russian word in a chemical paper will <u>almost always</u> have a certain specific English rendering that the danger is so great that in those exceptional cases where this word, for some reason or other, will have a different meaning, this exception will not be taken into account, yielding a meaningful but wrong translation.

In regard to the second method, the situation is even worse, and has lately become even more confused through the use of certain slogan terms like 'thesaurus' in this connection[6]. It is doubtless true that consideration of the immediate linguistic neighborhood of a given ambiguous word is a very powerful method, but it is again necessary to realize its limitations. I am not talking any more about those limitations which I pointed out through the use of my sample sentence. I am now talking rather about the fact that many MT workers seem to underestimate the importance of those cases of reduction of polysemy which cannot be obtained by looking at the immediate neighborhood, and even more so about the fact that partial successes in this direction have led many people to underestimate the depth of the remaining gap. Let me state rather dogmatically that there exists at this moment no method of reducing the polysemy of the, say, twenty words of an average Russian sentence in a scientific article below a remainder of, I would estimate, at least five or six words with multiple English renderings, which would not seriously endanger the quality of the machine output. It is looking at the quantities involved which creates a distorted picture with many people. Many tend to believe that by reducing the number of initially possible renderings of a twenty word Russian sentence from a million (which is the approximate number resulting from the assumption that each of the twenty Russian words has two renderings on the average) to some eighty (which would be the number of renderings on the assumption that sixteen words are uniquely rendered and four have three renderings apiece, forgetting now about all the other aspects such as change of word order, etc.) the main bulk of this kind of work has been achieved, the remainder requiring only some slight additional effort. We have before us another case of what, in a superficially different but intrinsically very similar situation, has been called the "80 per cent fallacy".[7] The remaining 20 per cent will require not one quarter of the effort spent for the first 80 per cent, but many, many times this effort, with a few per cent remaining beyond the reach of every conceivable effort.

NOTES

1) See "Decision procedures for structure in natural languages", Logique et Analyse, N.S., 2:19-29 (1959), which is a revised version of a talk given before the Colloque International de Logique, Louvain, September, 1958; the talk given before the Second International Congress on Cybernetics, Namur, September 1958, on "Some linguistic obstacles to machine translation" will be published in the Proceedings of this Congress.

2) See, especially, Chomsky, N., Syntactical structures, 's-Gravenhage,. Mouton & Co., 1957.

3) This memorandum is reprinted as Chapter 1 of Locke, W.N. and Booth, A.D., eds., Machine translation of languages, New York, John Wiley & Sons, 1955. The quoted passage appears there on page 21. For Reifler's and Karlan's studies, see ibid., p. 227.

4) See Dostert, L.E., "The Georgetown-I.B.M. Experiment", ibid., Chapter 8, especially pp. 129 ff.

5) I am afraid, therefore, that Weaver's hopes, reuttered in 1955 in his foreword to op. cit., that forthcoming research on logical syntax and semantics will make it possible for a computer to produce an output that would require no more than polishing up by a post-editor -- see ibid., p. VII -- will not materialize. I am singling out Weaver just because his misjudgment cannot be explained as being the result of vested interests.

6) Notice, e.g., that the very same -- fictitious! -- thesaurus approach that would correctly render 'pen' by 'plume' in the sentence 'The pen was in the inkstand' would incorrectly render 'pen' by 'plume' in the sentence 'The inkstand was in the pen'.

7) See Bull, W.E., Africa, Ch. and Teichroew, D., "Some problems of the 'word'", ibid., Chapter 5, p. 98.