Jeudi 3 septembre 2009 : Historicité et objectivité

B) Quelles périodisations pour l'histoire des idées linguistiques ?

Jacqueline Léon : *Technologie, demande sociale et traduction automatique : un cas de contraction de la périodisation dans l'histoire du récent*

Documents d'accompagnements :

1. *Translation*, Warren Weaver, 1949
2. *The present Status of Automatic Translation of Languages*, Yehoshua Bar-Hillel, 1960
3. *Language and Machines. Computers in translation and linguistics,* Automatic Language Processing Advisory Committee of the National Research Council, 1966

1 ♦ Translation*

WARREN WEAVER

There is no need to do more than mention the obvious fact that a multiplicity of languages impedes cultural interchange between the peoples of the earth, and is a serious deterrent to international understanding. The present memorandum, assuming the validity and importance of this fact, contains some comments and suggestions bearing on the possibility of contributing at least something to the solution of the world-wide translation problem through the use of electronic computers of great capacity, flexibility, and speed.

The suggestions of this memorandum will surely be incomplete and naïve, and may well be patently silly to an expert in the field—for the author is certainly not such.

A War Anecdote—Language Invariants

During the war a distinguished mathematician whom we will call *P*, an ex-German who had spent some time at the University of Istanbul and had learned Turkish there, told W. W. the following story.

A mathematical colleague, knowing that *P* had an amateur interest in cryptography, came to *P* one morning, stated that he had worked out a deciphering technique, and asked *P* to cook up some coded message on which he might try his scheme. *P* wrote out in Turkish a message containing about 100 words; simplified it by replacing the Turkish

* *Editors' Note:* This is the memorandum written by Warren Weaver on July 15, 1949. It is reprinted by his permission because it is a historical document for machine translation. When he sent it to some 200 of his acquaintances in various fields, it was literally the first suggestion that most had ever seen that language translation by computer techniques might be possible.

letters ç, ğ, ı̆, ö, ş, and ü by c, g, i, o, s, and u respectively; and then, using something more complicated than a simple substitution cipher, reduced the message to a column of five-digit numbers. The next day (and the time required is significant) the colleague brought his result back, and remarked that they had apparently not met with success. But the sequence of letters he reported, when properly broken up into words, and when mildly corrected (not enough correction being required really to bother anyone who knew the language well), turned out to be the original message in Turkish.

The most important point, at least for present purposes, is that the decoding was done by someone who did not know Turkish, and did not know that the message was in Turkish. One remembers, by contrast, the well-known instance in World War I when it took our cryptographic forces weeks or months to determine that a captured message was coded from Japanese; and then took them a relatively short time to decipher it, once they knew what the language was.

During the war, when the whole field of cryptography was so secret, it did not seem discreet to inquire concerning details of this story; but one could hardly avoid guessing that this process made use of frequencies of letters, letter combinations, intervals between letters and letter combinations, letter patterns, etc., *which are to some significant degree independent of the language used*. This at once leads one to suppose that, in the manifold instances in which man has invented and developed languages, there are certain invariant properties which are, again not precisely but to some statistically useful degree, common to all languages.

This may be, for all I know, a famous theorem of philology. Indeed the well-known *bow-wow, woof-woof,* etc. theories of Müller and others, for the origin of languages, would of course lead one to expect common features in all languages, due to their essentially similar mechanism of development. And, in any event, there are obvious reasons which make the supposition a likely one. All languages—at least all the ones under consideration here—were invented and developed by *men;* and all men, whether Bantu or Greek, Islandic or Peruvian, have essentially the same equipment to bring to bear on this problem. They have vocal organs capable of producing about the same set of sounds (with minor exceptions, such as the glottal click of the African native). Their brains are of the same general order of potential complexity. The elementary demands for language must have emerged in closely similar ways in different places and perhaps at different times. One would expect wide superficial differences; but

it seems very reasonable to expect that certain basic, and probably very nonobvious, aspects be common to all the developments. It is just a little like observing that trees differ very widely in many characteristics, and yet there are basic common characteristics—certain essential qualities of "tree-ness,"—that all trees share, whether they grow in Poland, or Ceylon, or Colombia. Furthermore (and this is the important point), a South American has, in general, no difficulty in recognizing that a Norwegian tree *is* a tree.

The idea of basic common elements in all languages later received support from a remark which the mathematician and logician Reichenbach made to W. W. Reichenbach also spent some time in Istanbul, and, like many of the German scholars who went there, he was perplexed and irritated by the Turkish language. The grammar of that language seemed to him so grotesque that eventually he was stimulated to study its logical structure. This, in turn, led him to become interested in the logical structure of the grammar of several other languages; and, quite unaware of W. W.'s interest in the subject, Reichenbach remarked, "I was amazed to discover that, for (apparently) widely varying languages, the basic logical structures have important common features." Reichenbach said he was publishing this, and would send the material to W. W.; but nothing has ever appeared.

One suspects that there is a great deal of evidence for this general viewpoint—at least bits of evidence appear spontaneously even to one who does not see the relevant literature. For example, a note in *Science,* about the research in comparative semantics of Erwin Reifler of the University of Washington, states that "the Chinese words for 'to shoot' and 'to dismiss' show a remarkable phonological and graphic agreement." This all seems very strange until one thinks of the two meanings of "to fire" in English. Is this only happenstance? How widespread are such correlations?

Translation and Computers

Having had considerable exposure to computer design problems during the war, and being aware of the speed, capacity, and logical flexibility possible in modern electronic computers, it was very natural for W. W. to think, several years ago, of the possibility that such computers be used for translation. On March 4, 1947, after having turned this idea over for a couple of years, W. W. wrote to Professor Norbert Wiener of Massachusetts Institute of Technology as follows:

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?

Professor Wiener, in a letter dated April 30, 1947, said in reply:

Second—as to the problem of mechanical translation, I frankly am afraid the boundaries of words in different languages are too vague and the emotional and international connotations are too extensive to make any quasimechanical translation scheme very hopeful. I will admit that basic English seems to indicate that we can go further than we have generally done in the mechanization of speech, but you must remember that in certain respects basic English is the reverse of mechanical and throws upon such words as *get* a burden which is much greater than most words carry in conventional English. At the present time, the mechanization of language, beyond such a stage as the design of photoelectric reading opportunities for the blind, seems very premature. . . .

To this, W. W. replied on May 9, 1947:

I am disappointed but not surprised by your comments on the translation problem. The difficulty you mention concerning Basic seems to me to have a rather easy answer. It is, of course, true that Basic puts multiple use on an action verb such as *get*. But, even so, the two-word combinations such as *get up, get over, get back*, etc., are, in Basic, not really very numerous. Suppose we take a vocabulary of 2,000 words, and admit for good measure all the two-word combinations as if they were single words. The vocabulary is still only four million: and that is not so formidable a number to a modern computer, is it?

Thus this attempt to interest Wiener, who seemed so ideally equipped to consider the problem, failed to produce any real result. This must in fact be accepted as exceedingly discouraging, for, if there

are any real possibilities, one would expect Wiener to be just the person to develop them.

The idea has, however, been seriously considered elsewhere. The first instance known to W. W., subsequent to his own notion about it, was described in a memorandum dated February 12, 1948, written by Dr. Andrew D. Booth who, in Professor J. D. Bernal's department in Birkbeck College, University of London, had been active in computer design and construction. Dr. Booth said:

A concluding example, of possible application of the electronic computer, is that of translating from one language into another. We have considered this problem in some detail, and it transpires that a machine of the type envisaged could perform this function without any modification in its design.

On May 25, 1948, W. W. visited Dr. Booth in his computer laboratory at Welwyn, London, and learned that Dr. Richens, Assistant Director of the Bureau of Plant Breeding and Genetics, and much concerned with the abstracting problem, had been interested with Dr. Booth in the translation problem. They had, at least at that time, not been concerned with the problem of multiple meaning, word order, idiom, etc., but only with the problem of mechanizing a dictionary. Their proposal then was that one first "sense" the letters of a word, and have the machine see whether or not its memory contains precisely the word in question. If so, the machine simply produces the translation (which is the rub; of course "the" translation doesn't exist) of this word. If this exact word is not contained in the memory, then the machine discards the last letter of the word, and tries over. If this fails, it discards another letter, and tries again. After it has found the largest initial combination of letters which *is* in the dictionary, it "looks up" the whole discarded portion in a special "grammatical annex" of the dictionary. Thus confronted by *running*, it might find *run* and then find out what the ending *(n)ing* does to *run*.

Thus their interest was, at least at that time, confined to the problem of the mechanization of a dictionary which in a reasonably efficient way would handle *all forms* of all words. W. W. has no more recent news of this affair.

Very recently the newspapers have carried stories of the use of one of the California computers as a translator. The published reports do not indicate much more than a word-into-word sort of translation, and there has been no indication, at least that W. W. has seen, of the proposed manner of handling the problems of multiple meaning, context, word order, etc.

This last-named attempt, or planned attempt, has already drawn forth inevitable scorn, Mr. Max Zeldner, in a letter to the *Herald Tribune* on June 13, 1949, stating that the most you could expect of a machine translation of the fifty-five Hebrew words which form the 23d Psalm would start out **Lord my shepherd no I will lack,** and would close **But good and kindness he will chase me all days of my life; and I shall rest in the house of Lord to length days.** Mr. Zeldner points out that a great Hebrew poet once said that translation "is like kissing your sweetheart through a veil."

It is, in fact, amply clear that a translation procedure that does little more than handle a one-to-one correspondence of words cannot hope to be useful for problems of *literary* translation, in which style is important, and in which the problems of idiom, multiple meanings, etc., are frequent.

Even this very restricted type of translation may, however, very well have important use. Large volumes of technical material might, for example, be usefully, even if not at all elegantly, handled this way. Technical writing is unfortunately not always straightforward and simple in style; but at least the problem of multiple meaning is enormously simpler. In mathematics, to take what is probably the easiest example, one can very nearly say that each word, within the general context of a mathematical article, has one and only one meaning.

The Future of Computer Translation

The foregoing remarks about computer translation schemes which have been reported do not, however, seem to W. W. to give an appropriately hopeful indication of what the future possibilities may be. Those possibilities should doubtless be indicated by persons who have special knowledge of languages and of their comparative anatomy. But again, at the risk of being foolishly naïve, it seems interesting to indicate four types of attack, on levels of increasing sophistication.

Meaning and Context

First, let us think of a way in which the problem of multiple meaning can, in principle at least, be solved. If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. "Fast" may mean "rapid"; or it may mean "motionless"; and there is no way of telling which.

But, if one lengthens the slit in the opaque mask, until 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.

The practical question is: "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?

This is a question concerning the statistical semantic character of language which could certainly be answered, at least in some interesting and perhaps in a useful way. Clearly N varies with the type of writing in question. It may be zero for an article known to be about a specific mathematical subject. It may be very low for chemistry, physics, engineering, etc. If N were equal to 5, and the article or book in question were on some sociological subject, would there be a probability of 0.95 that the choice of meaning would be correct 98% of the time? Doubtless not: but a statement of this sort could be made, and values of N could be determined that would meet given demands.

Ambiguity, moreover, attaches primarily to nouns, verbs, and adjectives; and actually (at least so I suppose) to relatively few nouns, verbs, and adjectives. Here again is a good subject for study concerning the statistical semantic character of languages. But one can imagine using a value of N that varies from word to word, is zero for *he, the,* etc., and needs to be large only rather occasionally. Or would it determine unique meaning in a satisfactory fraction of cases, to examine not the $2N$ adjacent *words,* but perhaps the $2N$ adjacent *nouns?* What choice of adjacent words maximizes the probability of correct choice of meaning, and at the same time leads to a small value of N?

Thus one is led to the concept of a translation process in which, in determining meaning for a word, account is taken of the immediate ($2N$ word) context. It would hardly be practical to do this by means of a generalized dictionary which contains all possible phases $2N + 1$ words long: for the number of such phases is horrifying, even to a modern electronic computer. But it does seem likely that some reasonable way could be found of using the micro context to settle the difficult cases of ambiguity.

Language and Logic

A more general basis for hoping that a computer could be designed which would cope with a useful part of the problem of translation is to be found in a theorem which was proved in 1943 by McCulloch and Pitts.[1] This theorem states that a robot (or a computer) constructed with regenerative loops of a certain formal character is capable of deducing any legitimate conclusion from a finite set of premises.

Now there are surely alogical elements in language (intuitive sense of style, emotional content, etc.) so that again one must be pessimistic about the problem of *literary* translation. But, insofar as written language is an expression of logical character, this theorem assures one that the problem is at least formally solvable.

Translation and Cryptography

Claude Shannon, of the Bell Telephone Laboratories, has recently published some remarkable work in the mathematical theory of communication.[2] This work all roots back to the statistical characteristics of the communication process. And it is at so basic a level of generality that it is not surprising that his theory includes the whole field of cryptography. During the war Shannon wrote a most important analysis of the whole cryptographic problem, and this work is, W. W. believes, also to appear soon, it having been declassified.

Probably only Shannon himself, at this stage, can be a good judge of the possibilities in this direction; but, as was expressed in W. W.'s original letter to Wiener, it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

This approach brings into the foreground an aspect of the matter that probably is absolutely basic—namely, the statistical character of the problem. "Perfect" translation is almost surely unattainable. Processes, which at stated confidence levels will produce a translation which contains only *X* per cent "error," are almost surely attainable.

And it is one of the chief purposes of this memorandum to emphasize that *statistical semantic* studies should be undertaken, as a necessary preliminary step.

The cryptographic-translation idea leads very naturally to, and is in fact a special case of, the fourth and most general suggestion: namely, that translation make deep use of language invariants.

Language and Invariants

Indeed, what seems to W. W. to be the most promising approach of all is one based on the ideas expressed on pages 16–17—that is to say, an approach that goes so deeply into the structure of languages as to come down to the level where they exhibit common traits.

Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers.

Thus may it be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication—the real but as yet undiscovered universal language—and then re-emerge by whatever particular route is convenient.

Such a program involves a presumably tremendous amount of work in the logical structure of languages before one would be ready for any mechanization. This must be very closely related to what Ogden and Richards have already done for English—and perhaps for French and Chinese. But it is along such general lines that it seems likely that the problem of translation can be attacked successfully. Such a program has the advantage that, whether or not it lead to a useful mechanization of the translation problem, it could not fail to shed much useful light on the general problem of communication.

REFERENCES

1. Warren S. McCulloch and Walter Pitts, *Bull. math. Biophys.*, no 5, pp. 115–133, 1943.

2. For a very simplified version, see "The Mathematics of Communication," by Warren Weaver, *Sci. Amer.*, vol. 181, no. 1, pp. 11–15, July 1949. Shannon's original papers, as published in the *Bell Syst. tech. J.*, and a longer and more detailed interpretation by W. W. are about to appear as a memoir on communication, published by the University of Illinois Press. A book by Shannon on this subject is also to appear soon. [A joint book, *The Mathematical Theory of Communication,* by Shannon and Weaver, was published by the University of Illinois Press in 1949—*Editors' Note*]

LANGUAGE AND MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the

Automatic Language Processing Advisory Committee

Division of Behavioral Sciences

National Academy of Sciences

National Research Council

Publication 1416

National Academy of Sciences National Research Council

AUTOMATIC LANGUAGE PROCESSING ADVISORY COMMITTEE

John R. Pierce, Bell Telephone Laboratories, Chairman
John B. Carroll, Harvard University
Eric P. Hamp, University of Chicago*
David G. Hays, The RAND Corporation
Charles F. Hockett, Cornell University†
Anthony G. Oettinger, Harvard University
Alan Perlis, Carnegie Institute of Technology

STAFF

A. Hood Roberts, Executive Secretary
Mrs. Sandra Ferony, Secretary

*Appointed February 1965

Contents

August 20, 1965

Dear Dr. Seitz:

In April of 1964 you formed an Automatic Language Processing Advisory Committee at the request of Dr. Leland Haworth, Director of the National Science Foundation, to advise the Department of Defense, the Central Intelligence Agency, and the National Science Foundation on research and development in the general field of mechanical translation of foreign languages. We quickly found that you were correct in stating that there are many strongly held but often conflicting opinions about the promise of machine translation and about what the most fruitful steps are that should be taken now.

In order to reach reasonable conclusions and to offer sensible advice we have found it necessary to learn from experts in a wide variety of fields (their names are listed in Appendix 20). We have informed ourselves concerning the needs for translation, considered the evaluation of translations, and compared the capabilities of machines and human beings in translation and in other language processing functions.

We found that what we heard led us all to the same conclusions, and the report which we are submitting herewith states our common views and recommendations. We believe that these can form the basis for useful changes in the support of research aimed at an increased understanding of a vitally important phenomenon—language, and development aimed at improved human translation, with an appropriate use of machine aids.

We are sorry that other obligations made it necessary for Charles F. Hockett, one of the original members of the Committee, to resign before the writing of our report. He nonetheless made valuable contributions to our work, which we wish to acknowledge.

Sincerely yours,

J. R. Pierce, Chairman
Automatic Language Processing
Advisory Committee

Dr. Frederick Seitz, President
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

July 27, 1966

Dear Dr. Seitz:

In connection with the report of the Automatic Language Processing Advisory Committee, National Research Council, which was reviewed by the Committee on Science and Public Policy on March 13, John R. Pierce, the chairman, was asked to prepare a brief statement of the support needs for computational linguistics, as distinct from automatic language translation. This request was prompted by a fear that the committee report, read in isolation, might result in termination of research support for computational linguistics as well as in the recommended reduction of support aimed at relatively short-term goals in translation.

Dr. Pierce's recommendation states in part as follows:

The computer has opened up to linguists a host of challenges, partial insights, and potentialities. We believe these can be aptly compared with the challenges, problems, and insights of particle physics. Certainly, language is second to no phenomenon in importance. And the tools of computational linguistics are considerably less costly than the multibillion-volt accelerators of particle physics. The new linguistics presents an attractive as well as an extremely important challenge.

There is every reason to believe that facing up to this challenge will ultimately lead to important contributions in many fields. A deeper knowledge of language could help:

1. To teach foreign languages more effectively.
2. To teach about the nature of language more effectively.
3. To use natural language more effectively in instruction and communication.
4. To enable us to engineer artificial languages for special purposes (e.g., pilot-to-control-tower languages).
5. To enable us to make meaningful psychological experiments in language use and in human communication and thought. Unless we know what language is we don't know what we must explain.
6. To use machines as aids in translation and in information retrieval.

However, the state of linguistics is such that excellent research that has value in itself is essential if linguistics is ultimately to make such contributions.

Such research must make use of computers. The data we must examine in order to find out about language is overwhelming both in quantity and in complexity. Computers give promise of helping us control the problems relating to the tremendous volume of data, and to a lesser extent the problems of data complexity. But we do not yet have good, easily used com-

Therefore, among the important kinds of research that need to be done and should be supported are (1) basic developmental research in computer methods for handling language, as tools to help the linguistic scientist discover and state his generalizations, and as tools to help check proposed generalizations against data; and (2) developmental research in methods to allow linguistic scientists to use computers to state in detail the complex kinds of theories (for example, grammars and theories of meaning) they produce, so that the theories can be checked in detail.

The most reasonable government source of support for research in computational linguistics is the National Science Foundation. How much support is needed? Some of the work must be done on a rather large scale, since small-scale experiments and work with miniature models of language have proved seriously deceptive in the past, and one can come to grips with real problems only above a certain scale of grammar size, dictionary size, and available corpus.

We estimate that work on a reasonably large scale can be supported in one institution for $600 or $700 thousand a year. We believe that work on this scale would be justified at four or five centers. Thus, an annual expenditure of $2.5 to $3 million seems reasonable for research. This figure is not intended to include support of work aimed at immediate practical applications of one sort or another.

This recommendation, which I understand has the endorsement of Dr. Pierce's committee, was also sent out for comment to the membership of the Committee on Science and Public Policy. While the Committee on Science and Public Policy has not considered the recommended program in computational linguistics in competition with other National Science Foundation programs, we do believe that Dr. Pierce's statement should be brought to the attention of the National Science Foundation as information necessary to put the report of the Advisory Committee in proper perspective.

Sincerely yours,

Harvey Brooks, Chairman
Committee on Science and Public Policy

Dr. Frederick Seitz, President
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

In computational linguistics and automatic language translation, we are witnessing dramatic applications of computers to the advance of science and knowledge. In this report, the Automatic Language Processing Advisory Committee of the National Research Council describes the state of development of these applications. It has thus performed an invaluable service for the entire scientific community.

Frederick Seitz, President
National Academy of Sciences

1. Aims and Methods, Survey and Critique

1.1 Introduction

Machine translation (MT) has become a multimillion dollar affair. It has been estimated[1] that in the United States alone something like one and one-half million dollars were spent in 1958 upon research more or less closely connected with MT, with approximately one hundred and fifty people, among them eighty with M.A., M.Sc. or higher degrees, working in the field, full or part time. 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. At a conference on MT that took place in Moscow in May 1958, 347 people from 79 institutions were reported to have participated. Not all participants need necessarily be actively involved in MT research. There exist two centers of research in MT in England, with a third in the process of formation, and one center in Italy. Outside these four 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 MIT, 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 the 1952 MT Conference I had prepared in mimeograph a survey of the state of the art [1]. 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. Time has come 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 the RAND Corporation [2], the other by Weik and Reitwiesner at the Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland [3]. 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 [4]. Brief histories of MT research are presented in the Introductory Comments by Professor Dostert to the Report of the Eighth Annual Round Table Conference on Linguistics and Language Study [5] as well as in the Historical Introduction to the recent book by Dr. Booth and associates [6].

The present survey is based upon personal visits during October and November 1958 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, upon talks with members of the two research groups in England, and upon replies to a circular letter 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, as well as, of course, upon a study of all available major publications including also, as much as possible, progress reports and memoranda; with regard to the USSR I had, unfortunately, to rely exclusively on available English translations of their publications and on reports which Professor Anthony G. Oettinger, of the Harvard Computation Laboratory, who had visited the major Russian research centers in MT in August 1958, was so kind to put at my disposal. Some of the purely technical information with regard to the composition of the various MT research groups, their addresses and budgets is presented in Appendix I in tabular form.

[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 cited in reference [3], say on p. 34 that "Dr. Panov's group consists of approximately 500 mathematicians, linguists and clerical personnel, all working on machine translation of foreign languages into Russian and translations between foreign languages with Russian as an interlanguage." No source for this figure is given, and it is likely that some mistake was made here.

Therefore, the innumerable specific advances of the various groups with regard to coding, transliterating, keypunching, displaying of output, etc., will be mentioned only rarely. But the list of references should contain sufficient indications for the direction of the reader interested in these aspects.

The order in which these groups will be discussed is: USA, Great Britain, USSR, others, following, with one exception, the order of degree of my personal acquaintance. Within each subdivision, the order will in general be that of seniority.

2.1 The USA Groups

2.1.1 THE SEATTLE GROUP

Professor Erwin Reifler of the University of Washington, Seattle, started his investigations into MT in 1949, under the impact of the famous memorandum by Weaver [17], and has since been working almost continuously on MT problems. The group he created has been constantly increasing in size and is at present one of the largest in the States. In February 1959, it published a 600-page report describing in detail its total research effort. This report has not reached me at the time of writing this survey (April 1959) which is the more unfortunate as the latest publication stemming from this group is a talk presented by Reifler in August 1957 [18], and I was, due to a personal mishap, unable to visit Seattle during my stay in the States. It is not impossible that my present discussion is considerably behind the actual developments.

The efforts of this group seem to have concentrated during the last years on the preparation of a very large Russian–English automatic dictionary containing approximately 200,000 so-called "operational entries" whose Russian part is probably composed of what was termed above (Section 1.3) "inflected forms" (as against the million or so inflected forms corresponding to the total Russian vocabulary of one hundred thousand canonical forms). This dictionary was to be put on a photoscopic memory device, developed by Telemeter-Magnetics Inc. for the USA Air Force, which combines a very large storage capacity with very low access time and apparently is to be used in combination with one of the large electronic computers of the IBM 709 or UNIVAC 1105 types. The output of this system would then be one version of what is known as *word-by-word translation,* whose exact form would depend on the specific content of the operational entries and the translation program. Both are unknown to me though probably given in the above mentioned report. Word-by-word Russian-to-English translation of scientific texts, if pushed to its limits, is known to enable an English reader who knows the respective field to understand, in general, at least the gist of the original text, though of course with an effort that is considerably larger than that required for reading a regular high quality translation, or else to enable an expert English post-editor to produce on its basis, with some very restricted use of the original text (in transliteration, if he does not know how to read Cyrillic characters), a translation which is of the same order of quality as that produced by a qualified human translator. However, no comparisons as to quality and cost between the Seattle MT system and human translation is given in the publications known to me. In any case, in view of the rather low quality of the machine output (word-by-word translation is theoretically a triviality, of course, though a lot of ingenuity is required to get the last drop out of it) the claim that the Seattle-Air Force system is "the most advanced translation system under construction" [19] is very misleading; even more misleading is the name given the photoscopic disc, "The USAF Automatic Language Translator Mark I" [20], which creates the impression of a special purpose device, which it is not.

The Seattle group started work towards getting better-than-word-by-word machine outputs in the customary direction of automatically changing the word order and reducing syntactical and lexical ambiguities (the Seattle group prefers to use the terms "grammatical" and "non-grammatical") but again little is known of actual achievements. One noticeable exception is Reifler's treatment of German compound words, which is an especially 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 will almost by necessity have to identify and analyze such compounds [21]. In the above mentioned 1957 talk, Reifler claimed to have "found moreover that only three matching procedures and four matching steps are necessary (sufficient?) to deal effectively with—that is, to machine translate correctly—any of these ten types of compounds of any[!] language in which they occur," [22]—a claim which sounds hardly believable, whose attempted substantiation is probably contained in the mentioned report. It is 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 [23].

For reasons given above, I must strongly disagree with Reifler's "belief that it will not be very long before the remaining linguistic problems in machine translation will be solved for a number of important languages" [24]. How dangerous such prophecies are is illustrated by another prophcy of Reifler's, to the effect that "in about two years (from August 1957) we shall have a device which will at one glance read a whole page and feed what it has read into a tape recorder and thus remove all human cooperation on the input side of the translation machines" [25]. The best estimates I am aware of at present mention five years as the time after which we are likely to have a reliable and versatile print reader (Section 1.3) at the present rate of research and development.

2.1.2 THE MIT GROUP

I started work on MT at the Research Laboratory of Electronics of MIT in May 1951. In July 1953, when I returned to Israel, Victor H. Yngve took over, steadily recruiting new assistants for his research. During the last years, the MIT group has laid great stress on its adherence to the ideal of FAHQT. For this purpose they regard 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 is rather uncertain in its belief that this prerequisite will be attained in the near future. In one of his latest publications, Yngve says: "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 into another without human intervention" [26]. 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. On the other hand, the MIT group believes, and I think rightfully, that the insights into the workings of language obtained by its research are valuable as such, and could at least partly be utilized in practical lower aimed machine translation by whomever 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 in this field [27].

The impact upon MT of Chomsky's recently attained insights into the structure of language is not quite clear. Since I presented my own views on this issue in a talk at the Colloque de Logique, Louvain, September 1958 [28], as well as in a talk given before the Second International Congress of Cybernetics, Namur, September 1958, a greatly revised version of which is reproduced in Appendix II, I shall mention here only one point. The MIT group believes, I think rightly, that Chomsky has succeeded in showing that the *phrase structure model* (certain variants of which are also known as *immediate constituent models*) which so far has 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 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 for a weak variant of this model, already 6 years ago [29] by a method that was later improved by Lambek [30] (cf. Appendix II)—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 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 and 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 cherishes similar doubts, and as a matter of fact my present evaluation derives directly from talks I had with him during my recent visit to the States.

The MIT group has, among other things, also developed a new program language called COMIT which, though specially adapted for MT purposes, is probably also of some more general importance [31], and whose use is envisaged also by other groups.[3] 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 realization of the enormous difficulties standing in the way of FAHQT. It is doubtful whether the development of a program language beyond some elementary limits is indeed necessary, or even helpful for more restricted goals. 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 FAHQT, will by necessity be led in the direction of studying learning machines.

[3] This information was given to me in a letter from Yngve.

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.

About the actual achievements of the MIT group with regard to MT proper little is known, apparently due to its reluctance to publish incomplete results. It is often felt that because of this reluctance other MT workers are wasting some of their time in treading over ground that might have already been adequately covered, though perhaps with negative results.

2.1.3 THE GU GROUP

The largest group working on MT in the States is that at Georgetown University, Washington, D.C., led by Professor Dostert. The GU Group comprises four subgroups. One of these is headed by Professor Garvin and has been engaged during the last two years exclusively in programming the mechanization 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 the details of their program debugged. They have produced a very large number of publications, in addition to a multitude of Seminar Work Papers of the Machine Translation Project of Georgetown University, of which I shall mention only two of the more recent ones [32, 33].

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. It is claimed that during the last few 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 above (Section 1.2) that far-reaching claims were made by one of the GU subgroups. This is the group headed by Miss Ariadne W. Lukjanow and using the so-called *Code Matching Technique* for the translation of Russian chemical texts. I expressed then my conviction that this group could not possibly have developed a method that is as fully automatic and of 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 of high quality and occasionally not even grammatical. True enough, this did not prevent the reader from understanding most of the time what was going on, but this would have been the case also for word-by-word translation, since the sample, perhaps due to its smallness, did not contain any of those constructions which would cause word-by-word translation to be very unsatisfactory. In contrast, however, with word-by-word translation which, if properly done, is hardly ever wrong, though mainly only because it is not real translation and leaves most of the responsibility to the post-editor, this translation contained one or two rather serious errors, as I was reliably told by someone who carefully went through the machine output and compared it with the Russian original. (I myself did not attend the demonstration, and my knowledge of Russian is rather restricted.)

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 semipublicly available document prepared by herself [34]. 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, in a single pass, 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 go through the source-language sentences in many passes, usually trying to isolate certain units first. I shall return to Miss Lukjanow's approach below (Section 2.1.9).

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 into 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 III, and that therefore the claim of the Lukjanow group is definitely false. There should be no difficulty for anyone who wishes to take the trouble 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 are in public 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 be of great help to everybody in the field. I understand that work on MT at Ramo-Wooldridge has been discontinued at the end of 1958, though perhaps only temporarily so.[6a]

2.1.6 THE HARVARD GROUP

The Harvard University group, headed by Professor Anthony G. Oettinger, stands in many respects quite apart from the others. First, it has busied itself for years almost exclusively with an exploration of the word-by-word translation method. Secondly, this preoccupation was accompanied by, and originated partly out of, a strong distrust of the achievements of other groups. Though it must be admitted that the possibilities of word-by-word translation from Russian into English have never before been so thoroughly explored as they were by this group, with many new insights gained, and that very valuable results were obtained as to the structure and construction of MT dictionaries, one may still wonder whether this group really struck the golden middle between utilizing other people's work in the field and distrusting their work, though there certainly were good reasons for the distrust on quite a few occasions.

The progress made by this group can be easily evaluated by comparing two doctoral theses submitted at Harvard University, the one—to my knowledge the first dissertation on MT—by Oettinger [40] in 1954, the other by Giuliano in January 1959 [41]. This second thesis seems to close an era and indicate the opening of a new one. The first five chapters describe the operation of the Harvard Automatic Dictionary, the methods for its compiling and updating, as well as a great variety of applications, in such thoroughness and detail that the impression is created that not much more is to be said on this subject. The last chapter, on the other hand, contains some interesting but tentative and almost untested remarks on what Giuliano calls a *Trial Translator* [42], i.e., an automatic programming system for the experimental production of better than word-by-word translations.

Out of the enormous amount of material contained in this thesis, let me dwell on those passages that are of immediate relevance to the question of the commercial feasibility of MT. The existing program at the Harvard Computation Laboratory can produce word-by-word Russian-to-English translations at a sustained rate of about 17 words per minute on a UNIVAC I, and about 25 words per minute on a UNIVAC II. This is 4–6 times more than an expert human translator can produce, but since UNIVAC II time is 100 times more expensive than a human translator's

[6a] *Note added in proof:* In the meantime, continuation of this project has been decided upon.

time, commercial MT is out of the question at present. Giuliano estimates that a combination of an IBM 709 (or UNIVAC 1105) with the photoscopic disc mentioned above (Section 2.1.1) would, after complete reprogramming—requiring some three programmer years—and a good amount of other development work, be able to produce translations at 20–40 times the present rate which, taking into account the increase in the cost of computer time, would still leave the cost of a word-by-word machine translation slightly above that of a high-quality human translation. The difference will, however, now be so slight that one may expect that any further improvement, in hardware and/or in programming, would reverse the cost relationship. This does not yet mean that true word-by-word MT will be in business. The cost of post-editing the word-by-word output in order to turn it into a passable translation of the ordinary type would probably be not much less than producing a translation of this quality without machine aid. As a matter of fact, senior research scientists having excellent command of scientific Russian and English, and extensive experience in technical writing, would be hampered rather than assisted by the automatic dictionary outputs in their present form.[7] The number of these individuals is, on the other hand, rather small and few of them can take the time from their scientific work to do a significant amount of translating and would have to be remunerated several times the ordinary professional translator's fee to be induced to spend more time on translating.

Altogether, it does not seem very likely that a nonsubsidized, commercial translation service will, in the next five years or so, find use for an automatic dictionary as its only mechanical device. However, as the Harvard group is quick to point out, an automatic dictionary is an extremely valuable research tool with a large number of possible applications, some of which have already proved their value. Let me add that in situations where speed is at a premium, high quality is not a necessary requisite, and human translators at a shortage for any price—such situations might arise, for instance, in military operations—automatic dictionaries would be useful as such for straight translation purposes.

The whole issue is, however, somewhat academic. There is no need to speculate what the commercial value of an automatic dictionary would be since the same computer-store combination that would put out a word-by-word translation can be programmed to put out better than word-by-word translations. This is, of course, the subject on which most MT groups, including the Harvard group itself as of this year, are working on right now. At what stage a winning machine-post-editor combina-

[7] This evaluation is taken from a paper by Giuliano and Oettinger, "Research on automatic translation at the Harvard Computation Laboratory," to be published.

Appendix I

MT Statistics as of April 1, 1959

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

Institution	Year of start of research	Number of workers	Full-time equivalents	Current yearly budget ($)	Project leader(s)
University of Washington Department of Far Eastern and Slavic Languages and Literature Seattle, Washington	1949	10 ?	6 ?	?	Erwin Reifler
Massachusetts Institute of Technology Research Laboratory of Electronics and Department of Modern Languages Cambridge 39, Massachusetts	1951	10 ?	6 ?	?	Victor H. Yngve
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
University of Michigan Willow Run Laboratories Ann Arbor, Michigan	1955	11	7	?	Andreas Koutsoudas
University of Pennsylvania Department of Linguistics Philadelphia, Pennsylvania	1956 ?	10 ?	3 ?	?	Zellig S. Harris
National Bureau of Standards Washington, D.C.	1958	3	2	25,000	Ida Rhodes
Wayne State University Department of Slavic Languages and Computation Laboratory Detroit, Michigan	1958	10	6	40,000	Harry H. Josselson Arvid W. Jacobson
University of California Computer Center Berkeley, California	1958	8	5	40,500	Louis G. Henyey Sydney M. Lamb
University of Texas Department of Germanic Languages Austin 12, Texas	1958	?	?	?	Winfred P. Lehmann

Sources primaires
- Bar-Hillel Y., 1953, “A Quasi-Arithmetical Notation for Syntactic Description” *Language* 29:47-58.
- Bar-Hillel Y., 1960, “The present Status of Automatic Translation of Languages” *Advances in Computers* vol.1, F.C. Alt ed. Academic Press, N.Y., London: 91-141.
- *Language and Machines. Computers in translation and linguistics*. 1966. A report by the Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences, National Research Council.
- Locke W.N and A.D. Booth (eds.), 1955, *Machine Translation of Languages,* 14 essays, MIT and John Wiley & son New York.
- Shannon Cl. E. et Weaver W., 1949, *The Mathematical Theory of Communication*, Urbana: University of Illinois Press.
- Wiener Norbert, 1948 & 1961, *Cybernetics : or the Control and Communication in the Animal and the Machine* MIT :the MIT Press.
- Yngve V.H., 1960 “A Model and an hypothesis for language structure”, *Proceedings of the American Philosophical Society*, vol.104, n°5:444-466
- Weaver W., 1955 [1949] *Translation* in *Machine Translation of Languages,* 14 essays, MIT et John Wiley :15-23
- Weaver Warren, 1970, *Scene of Change. A Lifetime in American Science*. New York : Charles Scribner’s Sons.

Sources secondaires
- Cori M. et Léon J., 2002, “La constitution du TAL. Etude historique des dénominations et des concepts”, *Traitement Automatique des Langues*, n°43-3 :21-55.
- Dahan Amy et Pestre Dominique (eds.), 2004, *Les sciences pour la guerre (1940-1960)* Paris : Editions de l’EHESS.
- Conway Flo & Siegelman Jim, 2005, *Dark Hero of the Information Age. In search of Norbert Wiener the Father of Cybernetics*, New York :Basic Books.
- Hutchins W.J., 1996, “ALPAC: The (In)famous Report”, *MT News International 4:* 9-12.
- Hutchins W.J., 2000, *Early Years in Machine Translation,* John Benjamins, Amsterdam, Philadelphia.
- Léon J., 1992, "De la traduction automatique à la linguistique computationnelle. Contribution à une chronologie des années 1959-1965", *Traitement Automatique des Langues,* vol.33, n° 1-2 :25-44.
- Léon J., 1999, “La mécanisation du dictionnaire dans les premières expériences de traduction automatique (1948-1960)”, *History of Linguistics 1996* vol. II (D.Cram, A. Linn, E. Nowak eds.) :331-340 John Benjamins Publishing Company.
- Léon Jacqueline. 2006. “La traduction automatique I: les premières tentatives jusqu’au rapport ALPAC” *Handbücher zur Sprach- und Kommunikationswissenschaft,* Sylvain Auroux, E.F.K. Koerner, Hans-Joseph Niederehe, Kees Versteegh (eds.) Walter de Gruyter and co., Berlin volume 3, Histoire des Sciences du Langage :2767-2774.
- Léon Jacqueline. 2006. “La traduction automatique II: développements récents” *Handbücher zur Sprach- und Kommunikationswissenschaft,* Sylvain Auroux, E.F.K. Koerner, Hans-Joseph Niederehe, Kees Versteegh (eds.) Walter de Gruyter and co., Berlin volume 3, Histoire des Sciences du Langage :2774-2780.
- Murray Stephen O. 1994, *Theory groups and the study of language in North America,* Studies in the History of the Language Sciences 69, John Benjamins Publishing company.

- Segal Jérôme, 2003, *Le Zéro et le Un. Histoire de la notion scientifique d'information au 20ème siècle*, Paris : Editions Syllepse.