OUTLINE OF THE RESEARCH

BY

ERWIN REIFLER

1. History of the Project

The history of machine translation (MT) research at the University of Washington in Seattle from its beginnings in November, 1949 until June 30, 1958 has already been outlined in our previous Technical Report.[1] This research, which was inspired by Dr. Warren Weaver's memorandum Translation[2] of July 15, 1949, and which was in 1952 and 1953 supported by grants from the Rockefeller Foundation, has since May, 1956 been sponsored by RADC, ARDC, USAF, by grants totaling $235,500.

Under the USAF contract, the University of Washington undertook to study the lexicographical, linguistic and engineering prerequisites for the automatic translation of scientific Russian into English. The University was specifically charged with the preparation of a Russian-English lexicon to be used in an automatic system designed by Dr. Gilbert W. King, then with the International Telemeter Corporation of Los Angeles and now with IBM. This automatic system includes a photoscopic memory device which has, or soon will have, sufficient storage capacity for all lexicographical requirements of machine translation.

Since Dr. King's automatic system did not yet include logical equipment for linguistic purposes, it was decided to attempt to solve as many bilingual linguistic problems as possible by purely lexicographical means. It was, however, clear from the outset that such a lexicographical solution of linguistic problems would be possible only in the case of certain types of divergent structure and multiple meaning. Therefore, investigations were initiated very early for the purpose of developing logical procedures and machine programs for the solution of the remaining linguistic problems.

During the initial, lexicographical, phase of the research, the project compiled a lexicon of about 14,000 Russian-English entries whose Russian "semantic units"[3] were taken from 111 Russian texts covering 40 fields of science. They did, of course, in these texts not occur in all their possible paradigmatic transformations. It was therefore necessary to supplement all those paradigmatic forms which did not happen to occur in the 111 texts, but were likely to occur in other scientific texts. This supplementation work had already been initiated during the previous report period. At the end of that period, in June, 1958, it was estimated that the supplementation would increase the number of entries to about 200,000.

The final count at the conclusion of the lexicographical phase of the project in June, 1959, showed that our estimate had been too generous. By that time the sponsors of the project had received from us a Russian-English MT lexicon of 170,563 entries on 556,141 IBM punch cards.

2. The Contents of the Lexicon

The lexicon contains a nearly complete collection of the general language vocabulary likely to occur in Russian scientific publications, and samples of Russian technical terms from forty fields of science. Not only individual free and bound forms, and not only the headwords of paradigmatic semantic units (infinitive of verbs, nominative singular of nouns, etc.) are here treated as lexical units, but also a number of bilingual idioms[4] and all relevant paradigmatic forms. This inclusion of bilingual idioms and of all relevant paradigmatic forms in the MT lexicon will contribute substantially towards a future conventionalization of the MT output.

3. The Predicted Output

An MT product based on the present form of our bilingual lexicon and of a quality as predicted by our sim-

[1] Erwin Reifler, "Outline of the Project," §1.0, Linguistic and Engineering Studies in the Automatic Translation of Scientific Russian into English. Seattle: The University of Washington Press, June, 1958.

[2] Warren Weaver, "Translation," Machine Translation of Languages, William N. Locke and A. Donald Booth, Cambridge: The Technology Press of the Massachusetts Institute of Technology, 1955.

[3] I.e., a single free or bound meaningful symbol, or free or bound symbol sequence, and any group of free symbol sequences which is idiomatic in terms of source-target semantics. Cf. also Erwin Reifler, "Some New MT Terms," the first paper in the Linguistic Analysis section of our previous report.

[4] Cf. Erwin Reifler, "MT Linguistics and MT Lexicography at the University of Washington," the second paper in the Linguistic Analysis section of this report, II/l.

ulated machine translations[5] can, of course, not at all compare with translations prepared by qualified human translators. It may, nevertheless, already be useful for purposes in which publication of the translation product is not desired or necessary. We do, however, by no means consider such an output as the final product, the ultimate goal of our research efforts, but only as an intermediate result, in fact only as a raw product whose inadequacies have to be studied and which has to undergo further refining processes. The reader is well advised to keep this fact, already stressed in our previous report,[6] in mind when he leafs through our simulated machine translations.

4. Syntactic Research

The investigations, mentioned earlier (§1.0), for the purpose of developing logical procedures and machine programs for the automatic resolution of source-target problems which cannot be solved by lexicographical means alone were continued after the completion of the lexicon in June, 1959. However, only a few months remained from that date until the time this report became due, and much of the intervening time had to be dedicated to the organization of the material and the preparation of this report. Consequently the syntactic research undertaken during this period is by no means complete and its results are not yet conclusive.

In languages like Russian the verbs are the form class which in most sentences occupy the most strategic position in the hierarchy of syntactic relationships, and, therefore, attempts to disentangle the net of these relationships best start with the predicate. During the short time still available for syntactic research, 240 verbs (each perfective verb and its imperfective correspondent counted as one) were examined in the light of certain relevant principles of transformational grammar, and their syntactic behaviour was then studied in order to determine their usefulness in the resolution of grammatical and nongrammatical ambiguities. After the elimination of all verbs with only one target correspondent, only 103 verbs (later further reduced to 97) remained, and the grammatical information they carry was then checked against some 2000 sentences in which they occurred. The details and preliminary results of this research are described by Dr. Micklesen in the third paper of the Linguistic Analysis section of this report.

5. Characteristics of the University of Washington Approach[7]

MT research at the University of Washington, has, it appears to me, hitherto differed mainly in three important respects from that of other research groups in this country and abroad. These are:

(1) The apparent emphasis on lexicography rather than on the structural linguistic aspects of the MT problem.
(2) The apparent emphasis on the language and vocabulary of many fields of science rather than of one field or subfield of science.
(3) The apparent emphasis on and preference for the "free-form approach" rather than what we may call the "stems and endings approach." This means the reception into the MT-operational lexicon not only of the undissected customary lexical forms such as the nominative singular of nouns, the present infinitive of verbs, etc., etc., but even of all relevant forms of the paradigms of all eligible semantic units of the source language.

The result of this "free-form approach" has been the MT-operational lexicon containing 170,563 Russian-English entries mentioned above. The Russian part of these more than 170,000 entries represents only about 14,000 Russian semantic units.

The decision to take such a course has by no means been an arbitrary one, but was based on the following facts:

(1) Our sponsors did not require us, during the initial phases of our project, to undertake structural analytic researches, but explicitly charged us to compile a Russian-English lexicon for MT purposes. In the course of the ensuing lexicographical studies I soon realized that in certain types of cases of higher frequency it is possible to solve grammatical and/or nongrammatical problems by lexicography alone—that is, without the necessity of logical procedures and logical machine operations. This approach I shall outline in section 6.0. We decided therefore to try to achieve an optimum of lexicography in which we would solve as many problems of grammatical and nongrammatical ambiguity as possible, leaving the unsolved problems to be dealt with by logical programs based on considerations of environment and structural hierarchy.
(2) As for our apparent emphasis on the language and vocabulary of many fields of science rather than of one field or subfield of science I have to stress the following two points:
    (a) The first is that we were privileged and fortunate to be asked by our sponsors to do our lexicographical and, later, linguistic work in consideration of a translation system which includes the truly ingenious photoscopic disc designed by Dr. Gilbert W. King of IBM. This permanent memory

[5] Cf. Appendix of the Linguistic Analysis section.

[6] Cf. op. cit. in footnote 1, §6.0.

[7] The following description has been presented in a paper read before the National Symposium on Machine Translation at the University of California, Los Angeles, February 2-5, 1960. It will be published in the Proceedings of the Symposium under the title, "Current MT Research at the University of Washington."

device has, or soon will have, a practically unlimited storage capacity so that the reduction of the size of the MT glossary has lost its initial significance. If specialized or idio-glossaries for one field or subfield of science will still be used in MT in the future, it will not be because of any limitations in the storage capacity of the permanent memory device.

(b) The second point is the fact that in scientific publications it is not the scientific and technical terms which present *a* serious semantic problem although they are frequently shared by more than one field of science and then sometimes require different translations. The important point I have to stress here is the fact that in scientific publications it is the nonscientific and nontechnical general-language vocabulary which constitutes the major semantic problem in our MT research. And this genera1-language vocabulary is shared by all fields and sub-fields of science and presents in all of them the same semantic problems. We decided, therefore, to compile the general-language vocabulary current in modern Russian scientific texts and to work out procedures for the pinpointing of intended meaning in the case of scientific and technical terms belonging to more than one field of science. This problem is presented in detail in the paper, "An Experiment in the Automatic Selection or Rejection of Technical Terms," by Dr. Lew R. Micklesen which forms part of the third paper (§1.2) in the Linguistic Analysis section of this report.[8]

These two reasons, namely the availability of Dr. King's photoscopic disc and the fact that the general-language vocabulary is shared by all fields of science, made it unnecessary for us to limit ourselves in our MT research to one single field or subfield of science.

(3) The reasons for our apparent preference for the "free-form approach" instead of the linguistically and engineeringwise really very attractive "stems-and-endings approach" I have frequently stated very emphatically. It is worthwhile repeating them here. The "stems-and-endings approach" means automatic dissection of input forms, automatic identification of the constituent stem and ending or endings, and the automatic determination of the intended meaning of the dissected free form in consideration of the information coded with the stem and the ending. All this requires, of course, quite a number of machine operations. Moreover, the dissection into stem and ending means the destruction or loss of the meaning or meanings the free form carries, and this lost semantic information has to be synthesized subsequently by the machine. All this is necessary as long as the storage capacity of the permanent memory is limited, and it may, for all I know, actually become the established practice of the MT of the future even when permanent memories with unlimited storage capacities are available. But since we at the University of Washington had the good fortune to be able to work in consideration of Dr. King's photoscopic disc, we felt that we should not increase our in any case already stupendous nongrammatical meaning problem by destroying the meaning or meanings of free forms through dissection into stem and ending.

We realized, however, from the very outset the importance of structural linguistics for MT. Already during the time of our concentration on lexicography, some members of the University of Washington MT project began with structural analytic researches for the purpose of elaborating logical procedures for the solution of all those problems lexicography alone cannot solve. The preliminary results have been published in our first comprehensive report. These researches were, as already stated above, continued after the completion of the MT-operational Russian-English lexicon and are reported in detail in the present report.

6. The Solution of MT Linguistic Problems Through Lexicography

I have already mentioned above that, in the course of lexicographical studies concerning the Russian-English MT-operational lexicon, I soon realized that in certain types of cases of higher frequency it is possible to solve grammatical and/or nongrammatical problems by lexicography and lexicographical procedures alone—that is, without the necessity of logical procedures and logical machine operations. The number of types whose source-target linguistic problems can be solved by purely lexicographical means is probably much larger than our present knowledge would lead us to believe. The following fact seems to indicate this. As stated earlier, it had been our original intention to achieve an optimum of lexicography in consideration of the logical limitations of Dr. King's automatic system and the large storage capacity of his permanent memory device. At the conclusion of the lexicographical phase I believed that we had attained this optimum by our emphasis on the genera1-language vocabulary of all fields of science, by our use of the "free form approach," by the supplementation of all relevant paradigmatic forms, by arrangements for the identification and translation of unpredictable compound forms, and especially by our treatment of genuine and pseudo-idioms.[9] Our previous report had, however, been hardly printed and distributed, when it occurred to me that I could vastly increase the number of instances in which an automatic system could be made to supply idiomatic translations without any logical operations and only on the basis of the information we are able to include in the bilingual lexicon. This approach I have described and exemplified in the paper, "MT Linguistics and MT Lexicography at the University of Washington."[10]

[8] Cf. also Udo K. Niehaus, "Automatic Pinpointing of Intended Meaning," in our previous report, Engineering Analysis, §8.0.

[9] Cf. previous report, "Outline of the Project," §2.2, 3.4, 4.1.

[10] The second paper of the Linguistic Analysis section of this report.

This unexpected development encourages me to believe in the possibility that future research may uncover additional lexicographical procedures of this kind, with the pleasant result of increasing the degree of agreement of the MT product with the idiomatic requirements of the target language without necessarily increasing the number *of* logical machine operations.

7. Conclusion

After all lexicographical means have been exhausted, there will still remain many unresolved problems of divergent source-target structure and of multiple meaning. Their resolution will depend on the success in the development of logical procedures and machine programs based on the results of syntactic research. It is possible that in these procedures and programs it will be necessary to consider again items of context already used in the lexicographical resolution of source-target problems. It is furthermore entirely possible that all source-target problems which can be solved by lexicographical means alone may in the future be solvable by logical procedures and programs.

There can, however, be no doubt that no logical programs can improve upon the readability of the output made possible in these cases by our lexicographical approach in which they are treated "as if they were idioms."