THE UNIVERSITY OF MICHIGAN
INDUSTRY PROGRAM OF THE COLLEGE OF ENGINEERING

TWO PAPERS

on

LANGUAGE TRANSLATION BY COMPUTERS

Andreas Koutsoudas
Robert R. Korfhage

December, 1956

IP-196

ACKNOWLEDGEMENT

We wish to express our appreciation to the Engineering Research Institute for permission to give this paper limited distribution under the Industry Program of the College of Engineering.

The research conducted on this subject was directed toward accomplishment of a Tri-Service Contract (DA-36-039-SC-52654) administered by the Army Signal Corps.

TABLE OF CONTENTS

MECHANICAL TRANSLATION

In view of the recent advances of the U.S.S.R. in both the scientific and technological spheres, it has become highly important that U.S. scientists and engineers have available the best possible data on Russian developments. At the present time such data is largely unavailable due to the mass of material to be translated, the inherent slowness of the translation and the fact that a relatively small number of technical and scientific translators are available. This being the case, several individuals in this country and abroad have seriously investigated the possibility of language translation by use of modern high-speed electronic computers and other data processing equipment.

The first decade of these investigations will end this year. Until 1952 all work was done by small groups of men who had little communication with each other. But in the spring of that year, the group of men at M.I.T who were interested in this problem called a four-day international conference on the subject. This conference, attended by the eighteen men who were then working on mechanical translation, did much to coordinate the work being done by the various groups in this country and England.

In 1954 I.B.M. and Georgetown University announced that they had successfully programmed a large-scale computer to do translation from Russian into English. Although the program which they had developed was a small demonstration problem which could not very well be expanded into a working translation program, the experiment at least showed that it is possible for a computer to do such work.

Enough work had been done by 1954 that some better means of communication between those interested had to be found. As a result, the M.I.T. group began publication of MT (Mechanical Translation). This journal has appeared at irregular intervals since then, as material has become available.

One of the most significant contributions to the field was the publication in 1955 of the book Mechanical Translation of Languages by W. N. Locke and A. D. Booth. This book, the only one which has been written on the subject, contains most of the major previously unpublished results in the field.

In the winter of that year, a study was begun at the University of Michigan into the possibility of using MIDAC, the Michigan Digital Automatic Computer, to translate Russian scientific texts into English. A thorough study of the available literature on mechanical translation convinced us that existing techniques were not the best possible, and that hence we would have to undertake a series of linguistic studies before formulating the computer program.

Clearly the simplest possible translation is word for word. To achieve such a translation, one would merely have to establish a dictionary assigning to each Russian word the corresponding English one. However, several problems arise.

Perhaps the most obvious problem is that of word order. A priori, there is no reason to suppose that Russians say things in the same order that English speaking people do. And indeed, using a word for word translation one would soon uncover such phrases as "not are" for "are not". While this particular case is not very troublesome, it is conceivable that a change in word order could alter the meaning, as in "man bites dog" and "dog bites man". Fortunately, studies by K. E. Harper of U.C.L.A. indicate the the sentence structure of technical Russian is quite similar to that of technical English, and that therefore the problem of word order is a minor one.

Another problem which readily comes to mind is that of actual word forms. Russian is a highly inflected language, with a total of fifty-eight orthographically distinct endings, of which as many as twenty-nine may apply to any given word. Clearly then, to list each word in all of its forms would greatly enlarge the dictionary. Again, however, the situation is not impossible. Our studies confirm Harper's findings here also, namely that since English has very few inflectional endings, most of the Russian endings are inconsequential in translation.

Thus we are left with two reasonable procedures: either to list a word in all forms with significant endings and also list the stem of the word for the other cases, or to list merely the stem of the word and treat all endings separately. We have chosen a combination of these methods with heavy emphasis on the latter.

One other problem occurs. It is not in general true that a given word in one language will have exactly one correspondent in another language. Thus we must be able to choose the proper meaning for a Russian word whenever it occurs in a given text. The solution to this problem depends, as we will see, on both the grammatical ending associated with this particular use of the word (plurals, etc.), and on the context in which the word occurs.

The starting point of any translation program is a dictionary. As a beginning we have restricted our studies to theoretical and experimental physics. From works in this field we selected approximately 64,000 words of text as an initial basis for our dictionary. This has given us a dictionary of some 3,000 words.

We do not know yet the ultimate size of our dictionary. One limitation is imposed by the computer we use. MIDAC currently has a high-speed storage of 512 "words" and a magnetic drum storage of 6,144 "words". The latter is being enlarged to 36,858 "words". Each "word" consists of forty-five binary digits, enough to represent seven letters, with three digits left over. These extra digits are used as hyphens in words of more than seven letters, and for various other purposes. With this storage we expect to be able to use a dictionary of around 7,000 words.

With a dictionary of this size the problem of arranging the words in an efficient order assumes importance. Certainly it is not desirable to have to search several thousand words to find a commonly used word. Rather, it would seem advantageous to place the frequently used words at the beginning of the dictionary, where the machine will find them quickly. A word count on a sample such as our 64,000 words will give a reasonable approximation to the relative frequencies of occurrence of words in our restricted segment of the language, so that such an arrangement for our dictionary would be simple to construct. In general, if the sample of the language is well chosen, words which are not encountered will occur infrequently, and may be added, naturally, at the end of the dictionary.

The difficulty with this system is that it only apparently solves the problem, for although the words near the end of the dictionary individually occur quite rarely, in the aggregate they form a large portion of any piece of text. A possible alternative to this system is an alphabetically arranged dictionary. However, in this case it is quite difficult to add words to the dictionary.

The best solution to the problem of dictionary arrangement seems to be a combination of the above. That is, it seems desirable to arrange the dictionary alphabetically on the first two or three letters of the words, and then to arrange each of the resulting groups of words by their relative frequency. Such an arrangement would give better search times than a straight alphabetic ordering, and would also eliminate the long search for rarely used words necessary with a purely frequency-arranged dictionary. The best point at which to change from one arrangement to the other is not known, and will not be known until we have been able to compare translation times with a given dictionary, arranged in various possible ways.

The other important ingredient of any translation program is the set of rules by which the machine will operate on the text which it is to translate. In our preliminary work we have divided the rules into two subsets - one group dealing with word order and endings, and the other for handling multiple meanings. We found that much of the information conveyed by the inflectional endings is also carried by the order in which the words occur. For example, in technical Russian the occurrence of two nouns together means, with one exception, that the second is in the genitive case and must, in English, be preceded by "of". Thus we have been able to eliminate consideration of most of the inflectional endings, and to establish about a dozen rules which deal mainly with word order and give us most of the syntactical and grammatical information necessary. This permits us to list in the dictionary merely the stem or root of most words. In the few cases where our rules do not apply we have found it necessary to list two or three different forms of the word.

The solution to the multiple meaning problem is more complex. The problem is simplified somewhat by restricting our language to technical fields, but there still remain ambiguities. It seems reasonable to suppose that in any given

usage the correct meaning of a word (among several) can be determined from the context, that is, from the surrounding words. In practice we have never found it necessary to consider more than six words, and almost always two or three suffice. Basically then the problem consists of recognizing such contextual sequences and arranging the meanings of the words so that the machine will select the correct ones. The latter part of this is complicated by the fact that the same words may occur in different order and have different meanings, as in the "man bites dog" case. We have been unable to find an ordering of meanings which will give the correct meaning in all cases, although we have been successful ninety to ninety-five per cent of the time.

One phase of the multiple-meaning problem which deserves special attention is the handling of idioms. Some of the early investigators, notably Y. Bar-Hillel of the Hebrew University in Jersualem, proposed that a separate idiom dictionary be created, and that all words be first checked for possible occurrence in this dictionary. We felt that since the use of idioms is relatively limited (one or two per page of text) such a procedure is rather wasteful of computer time, and that it would be better to consider an idiom as merely another possible meaning of the words comprising it. In idioms consisting of three or more words, it was found that one of the central words generally had only one or two meanings other than the idiomatic one, whereas the first and last words usually had many possible English equivalents. Thus it seemed natural to attach the whole meaning of the idiom to a central word and to assign no English equivalent at all to the other words when they occurred in an idiom. With two-word idioms such a scheme is not always possible, and they are handled more nearly as are other multiple-meaning cases.

Our results to date are very encouraging, but far from perfection. We have a system which will yield a readable translation of a Russian physics paper, and will do this without the necessity of someone editing the Russian text before it is given to the machine or the English translation produced. The result will not be of literary quality. There will be grammatical irregularities, misspellings (arising from the fact that we transliterate words, such as proper names, which are not in the dictionary), and ambiguous passages. However, the result will be good enough that an American physicist will be able to make sense of what he reads.

There are many improvements which can be made in our program. Certainly the extension of the translation scheme to other fields and to other languages is desirable. Other groups are working on this, particularly on the translation of German. Such an extension would require new dictionaries and a revision and expansion of the set of rules. Translations from other languages would undoubtedly require completely new rules.

Even without expanding our program to other fields there is much work to be done on the dictionary and rules. One of the most important unsolved problems is the use of the definite and indefinite articles. Russian has no articles, and hence we must add them where they would occur in English. However, the decision on which, if either, to use before a given noun seems highly complex. Undoubtedly the translation program which will be in use a few years from now will be very different from the rather simple one we have at present.

Improvements are being made too in the computers and data processors which can be used for mechanical translation. Machines are being built which will be faster than existing computers and will have larger memories. Currently the given text must be typed into the computer by a human operator. Devices are being designed which will enable the machine to literally read a printed page, thus eliminating the inherently slow typing process.

The problem until now has been to provide some translation - any translation - of papers and books which have been unavailable to American scientists and engineers because of a language barrier. Now that the way has been opened, the problem has become that of improving the quality of the translation and of increasing the diversity of material which can be handled. Within the next few years, the mechanical translation of languages will become one of the major tools used by our scientists and engineers.

MECHANICAL TRANSLATION AND THE PROBLEM OF MULTIPLE MEANING

The University of Michigan undertook research, late last fall, in the analysis of language structure for mechanical translation. Emphasis was placed on the use of the contextual structure of the sentence as a means of reducing ambiguity and on the formulation of a set of operative rules which an electronic computer could use for automatically translating Russian texts into English. This is a preliminary report on the latter phase of the problem, stating the results and suggesting a practical method for handling idioms and the problem of multiple meanings.

It was decided that the first work would be done on Russian texts in physics, both because of the interest in this field and because of the general availability of texts.[1] If this work proves successful, it will form a basis for work in other scientific, technical, and military fields.

A text was selected from a Russian journal on experimental and theoretical physics.[2] It was chosen to present most of the expected difficulties; i.e. stylistic, orthographical, grammatical, etc. On the basis of this text a vocabulary was set up and fifteen rules were established. It should be realized, of course, that neither the vocabulary nor the rules were in generally applicable form. The vocabulary was simplified by applying a "one form, one meaning" rule whenever possible. Thus, inflectional endings were stripped from most word stems although in some cases a word was listed with two or three specific endings. Most words were given their scientific meaning only. Some words, however, occurred in more than one sense, or were combined with others to form idioms; in which case more than one meaning had to be listed. Finally, the words were listed in conventional grammatical categories; i.e., verb, noun, adjective, etc.

In the long run, we expect that the concept of conventional categories will be completely abandoned. What we hope to have, instead, are word groups; the interaction of which will provide the grammatical and syntactical information needed.[3]

---

[1]Some work has already been done in this field. See K. E. Harper, "A Preliminary Study of Russian". Machine Translation of Languages, The Technology Press of the Massachusetts Institute of Technology and John Wiley & Sons, Inc., New York, 1955.

[2]Zhurnal Eksperimental'noi I Teoretichesk'oi Fiziki Vol. 26, No. 2, pp. 189-207, Feb. 1955

[3]The need for such grouping has been made apparent by V. H. Yngve in his article "Sentence for Sentence Translation", MT Vol. 2 No. 2, Nov., 1955

The rules were developed empiricially by analysis of the essential processes undertaken by a human mind in translating a foreign text. It was found that most of the rules involved either word order or the grammatical functions which in Russian are indicated only by case endings and which in English might be classified by inserting a preposition. In most cases the rules concerning word order were sufficient to eliminate the necessity of referring to endings. To test the adequacy of the rules, several volunteers who had no knowledge of Russian were asked to translate the original text, using only our rules and vocabulary. Except for random, minor stylistic faults, it turned out that the resulting translations were clear and accurate. Being convinced that the rules are as complete as is practicable for the text, we are currently enlarging the vocabulary in preparation for future tests on different texts.

Perhaps the most significant result thus far is the success in handling multiple meanings, which has given us an insight into the problem of idioms. Although the problem of ambiguity as exemplified by this situation was greatly reduced by the use of a highly specialized vocabulary, the situation still occurred and a means for solving it had to be found. Published results on this problem have, generally, involved either a post-editor or a separate idiom dictionary.[4] These methods seem undesirable particularly in view of the additional computer time required for translation. Consequently, a method was developed which, it is felt, is highly applicable. The assumption was made that the specific meaning of a word could be determined from its context. It developed that not only is this assumption valid, but in fact we need not consider sequences of more than four words. The method used is the following:

All possible meanings of a word are listed, consecutively, in the order (1), (2), ...(n). In general, in order to have corresponding meanings mesh, it will be necessary to list some meanings for each word more than once, and to include some blank translations. When a word with multiple meanings is encountered, the number (n) of meanings is noted and translation is postponed. Subsequent words are examined for the number of possible meanings of each, until a word (X) with a single meaning is encountered. If there is only one word in the sequence preceding X, then the first listed meaning is assigned to this word. If there is more than one word in the sequence preceding X, we determine (M), the minimum of all (n) noted in the sequence. Let us denote by (i) [A] the i-th meaning of a word A, and by $\underline{\times}$ a blank (null) translation.

Given a two-word sequence, A B, we consider (M)[A] and (M)[B]. If neither of these are blank, we translate, assigning meaning (M) to each word. If either of these is blank, we consider (M-1)[A] and (M-1)[B] , and apply the same test to these. In this way we find the highest numbered meaning which is not blank for either A or B and assign this meaning to each.

---

[4] In reference to these methods, see, for example: "The Treatment of Idioms" by Y. Bar-Hillel, typewritten, 8 pages; "A Study for the Design of an Automatic Dictionary" by A. G. Oettinger, doctoral thesis, Harvard University, '54.

Given a three-word sequence, A B C, we consider (M)[B]. If (M)[B] is X we consider successively meanings M-1, M-2,...., as above, and assign finally to all three words the highest numbered meanings which is non-blank for all. If (M)[B] is not X, then if (M)[A] and (M)[C] are both X we assign meaning (M) to the three words; otherwise we search meanings M-1, M-2, ... of all three words, applying the above rule.

In a four-word sequence, ABCD, (M)[B] is again considered. The procedure followed is that used for a three-word sequence, except that (M)[D] must be considered along with (M)[A] and (M)[C].

In all cases, if no translation is found by the above procedure, we assign to each word meaning (1).

By properly ordering the meanings for each word (listing some meanings several times if necessary) it has been found possible to obtain valid translations for over 96% of the two-word sequences[5] and for over 90% of the three-word sequences which might occur.[6] It is not known how the difficulties in "properly" ordering the meanings will multiply as the vocabulary is increased. With each new word (or meaning) added, the order of meanings previously listed may have to be changed so as to maintain consistency as much as possible.

In this system an idiom is handled as merely an additional meaning which is possible. A study of the structure of three-word idioms showed that generally the second word had the least number of meanings. On this basis it was decided to assign to the second word the entire idiomatic meaning, and to supply corresponding X translations for the other two words. Thus, for example, the Russian idiom ПО СУТИ ДЕЛА ("actually") would appear as ПО= X , СУТ = actually, ДЕЛ = X . (Note the dropped inflectional endings.)

To illustrate this method, let us consider the eight Russian words
ТОМ, ДЕЛ, СУТ, ЦЕЛ, ТЕОРИ, Ь, О and ПО
From these eight words it is possible to form 56 two-word sequences and 336 three-word sequences. However, of these only 29 two-word and 106 three-word sequences are linguistically possible.[7] By working with these 135 sequences

---

[5] The two exceptions which occurred, ПО ДЕЛУ and ЦЕЛЬ easily handled by separately listing ДЕЛ in the form ДЕЛУ and ЦЕЛ in the form ЦЕЛЬ

[6] These figures are based on the possible sequences without reference to their relative frequency of occurrence in actual use.

[7] It is assumed, of course, that the appropriate inflectional endings are supplied in each case. The list of sequences, with translations, is available on request.

it was found that the arrangement of meaning given in Table I is the best possible. There seem to be no algorithms for ordering the meanings, other than that the idiomatic meaning, if any, by the last meaning listed for at least one of the words.

TABLE I

ТОМ	ДЕЛ	СУТ	ТЕОРИ	ЦЕЛ	Ь	ПО	О
1. that 2. X	1. fact 2. X	1. essence 2. actually	1. theory 2. X 3. theory	1. purpose 2. X 3. X 4. order to 5. target	1. in 2. X 3. in 4. in 5. X	1. by 2. X 3. according to 4. X 5. at	1. about 2. X

It may be noted that on the basis of only the three words ПО, СУТ, ДЕЛ, and the shorter arrangement of meanings given in Table II suffices.

TABLE II

ПО	СУТ	ДЕЛ
(1) by	(1) essence	(1) fact
(2) X	(2) actually	(2) X

It will be observed that there is a certain amount of redundancy inherent in this system. However, it is felt that this is a minor fault; first, because the percentage of redundant meanings in the entire vocabulary appears to be small (around five per cent) and second, because this plan does not require a separate idiom dictionary or other special devices which tend to increase computer translation time. Although further research is necessary for the complete development of this method, we believe that the theory used is valid and that it eventually will lead us to the solution of most multiple-meaning problems.