MECHANICAL TRANSLATION

DEVOTED TO THE TRANSLATION OF LANGUAGES WITH THE AID OF MACHINES

VOLUME FOUR, NUMBER THREE DECEMBER, NINETEEN FIFTY-SEVEN

News

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

The establishment of a Center for Communication Sciences at the Massachusetts Institute of Technology was announced on May 10, 1958 by Dr. Julius A. Stratton, Acting President. The new center, which will be concerned chiefly with basic research having no direct military application, will be under the direction of a steering committee composed of Dr. Jerome B. Wiesner, Dr. Claude E. Shannon, Dr. Gordon S. Brown, Dr. Robert M. Fano, Dr. Roman Jakobson, and Dr. Walter A. Rosenblith.

Studies of communication functions of the nervous system and of such machines as computers as well as studies of communication between the two will be conducted in the new center by a group of scientists and engineers. Among the experimental studies that have been carried out by people in the Communication Sciences group are those dealing with: translation of languages by machine, electronic devices for aiding the blind and deaf, synthesis of human speech, compression of speech, and analysis of electrical activity of the brain by electronic computers.

* * *

Advanced research in problems of German syntax will be conducted at M.I. T. during the summer program in MT. In addition to the regular members of the M.I.T. group, other participants in the program are: B. Ulvestad, University of Bergen; L. Brandwood, Birkbeck College, London; S. Werbow, University of Texas; J. Gough, Georgia Institute of Technology; D. Dinneen, University of Kansas; and S. Rogovin, Columbia University.

CONFERENCE IN MOSCOW

On May 15-21, 1958 a Conference on Mechanical Translation was held in Moscow. Some seventy papers were presented by most of the Soviet linguists and computer experts who have worked with MT. The problems discussed ranged from linguistic theory to algorithms for mechanical translation from particular languages. Reports are printed in "Tezisi Konferencii po Mashinnomu Perevodu," printed by the Ministerstvo Vyshego Obrazovanija SSSR, 1-i Moskovskii gosudarstvennyi pedagogucheskii institut inostrannyh jazykov, 1958.

TEXTS AVAILABLE

100,000 words of English texts and 100,000 words of German texts are available in machinable form. In both cases the texts came from newspaper sources and represent news items, feature columns, and stories. All punctuation marks, carriage returns, and special symbols have been retained. The texts are available in any one of several formats and either on punched cards or magnetic tape for use in an IBM type 704 computer. Those who have need for texts such as these for research purposes may obtain copies by making arrangements with the undersigned.

Victor H. Yngve
Room 20B-101D
M. I. T.
Cambridge, Mass.

Some New Terminology

Erwin Reifler, University of Washington, Seattle, Washington

MT research requires cooperation between engineers and linguists. It is important, therefore, to develop a uniform linguistic terminology that can be understood and used by engineers. Furthermore, it is necessary that linguists develop an understanding of the engineering problems involved. The results of cooperation between linguists and engineers working with the MT Pilot Model at the University of Washington are presented here.

THE LINGUIST interested in pioneering in MT has to struggle with two difficult problems from the very outset: 1) the formulation of an adequate linguistic terminology that can be understood and used by the engineer, and 2) an understanding of the engineering problems involved. During our eight years of MT research at the University of Washington we have had the great advantage of close cooperation between linguists and engineers. I wish to submit for discussion under the heading of "Terminology" some of the results of this cooperation.

Recent developments in MT research at the University of Washington have necessitated the redefinition of some old linguistic terms and the formulation of some new ones. They concern the concepts of MT symbols, i.e., all graphic symbols used in the machine translation process. These MT symbols consist of the Control Symbols and Contextual Symbols.

1. Control Symbols — MT symbols which, coded into the machine memory, control certain steps in the translation process. Since they are not contextual symbols, they appear neither in the input nor in the output.

2. Contextual Symbols — the minimal contextual constituents used to produce a material stimulus for a machine-operational step relevant for MT, such as an alphabetic letter, a numerical figure, a dollar sign, a punctuation mark, a single space. Contextual symbols consist of Input Symbols and Output Symbols.

3. Input Symbols include all contextual symbols that may appear in a source text.

4. Output Symbols include:
   a) Letter symbols of the target alphabet
   b) Symbols for the numerals
   c) Punctuation symbols
   d) Editing symbols — target symbols intended to aid in the interpretation of the MT product. Examples are subscript numbers which are attached to some target equivalents to pinpoint the field or fields of science to which the scientific meanings of certain semantic units of the source language belong. (The term "semantic unit" will be explained below.)

5. Free Symbol — a contextual symbol preceded and followed by space. It is always meaningful and always used to symbolize both grammatical and non-grammatical meaning. An example is English 'I'.

6. Bound Symbol — a contextual symbol either not preceded or not followed, or neither preceded nor followed by space. We distinguish
   a) Left-bound symbols
   b) Right-bound symbols
   c) Twice-bound symbols

7. Meaningful Bound Symbol — a contextual symbol used to symbolize:
   a) Grammatical meaning, i.e., left-bound "s" in "father's, fathers", the right-bound " ' " in " 's" which indicates that the following "s" is a substantive ending, the twice-bound "o" in "arterio-sclerosis."

b) Non-grammatical meaning, i.e.., the left-bound "g" which distinguishes the meaning of "pang" from that of "pan", the right-bound "s" which distinguishes the meaning of "span" from that of "pan", the twice-bound "a" distinguishing the meaning of "seat" from that of "set."

c) Both grammatical and non-grammatical meaning, i.e., right-bound "o" distinguishing the grammatical and non-grammatical meaning of описать 'describe' (perfective aspect) from that of писать 'write' (imperfective aspect), left-bound "я" distinguishing the grammatical and non-grammatical meaning of ломя 'breaking' from that of лом 'crowbar', twice-bound "ж" distinguishing the grammatical and non-grammatical meaning of между 'between' from that of меду ' of the honey'.

8. Meaningless Bound Symbol — a bound symbol not intended by the author of a source text to symbolize anything, but treated as a separate entry by the MT planners in order to overcome engineering difficulties due to certain limitations of the MT equipment. An English example is the arbitrary left-bound final symbol "n" in "misinterpretation" which consists of 17 letters. If, for example, the input equipment cannot handle free symbol sequences longer than 16 letters, then "misinterpretation" may be split arbitarily into two constituents, the first of which contains the first 16 letters while the second consists of only one letter. These two constituents would then form two separate entries in the machine memory.

9). Symbol Sequence — a sequence of contextual symbols not interrupted by space.

10. Free Symbol Sequence — a symbol sequence preceded and followed by space. A free symbol sequence is always meaningful and is always used to symbolize both grammatical and non-grammatical meaning.

11. Bound Symbol Sequence — a symbol sequence either not preceded, or not followed, or neither preceded nor followed, by space. We distinguish:

a) Left-bound symbol sequence
b) Right-bound symbol sequence
c) Twice-bound symbol sequence

12. Meaningful Bound Symbol Sequence — a bound symbol sequence used to symbolize:

a) Grammatical meaning, i.e., left-bound "ren" in "children", and right-bound "be" in "befall" which changes the intransitive meaning of "to fall" into a transitive meaning, twice-bound ыв distinguishing the grammatical meaning of описывать 'to describe' (imperfective aspect) from that of описать 'to describe' (perfective aspect).

b) Non-grammatical meaning, i.e., left-bound "et" distinguishing the meaning of "ballet" from that of "ball", right-bound "bl" distinguish ing the meaning of "bleat" from that of "eat", twice-bound "ur" distinguishing the meaning of "gourd" from that of "god".

c) Both grammatical and non-grammatical meaning, i.e., left-bound "shore" in "seashore", right-bound "sea" in "seashore", and twice-bound "en" in "disentomb".

13. Meaningless Bound Symbol Sequence — a bound sequence not intended by the author of a source text to symbolize anything, but treated as an individual entry by the MT planners in order to overcome engineering difficulties due to certain limitations of the MT equipment. An English example is the meaningless left-bound symbol sequence "ss" in "irreconcilableness" which consists of 18 letters. The MT planners would have to split this free symbol sequence into two arbitrary constituents containing 16 and 2 letters respectively, and enter them as separate entries into the machine memory if the available input equipment cannot handle free symbol sequences longer than 16 letters.

14. Group of Free Symbol Sequences — a complete text or any part of a text, chapter, section, sentence or clause consisting of two or more free symbol sequences which symbolize a meaning intended by the author of the source text.

15. A Semantic Unit — a single free or bound meaningful symbol or symbol sequence, and any group of free symbol sequences which is idiomatic in terms of source-target semantics.

With the growth of MT development and the increase in the number of MT pioneers it is becoming more and more important to achieve some uniformity in linguistic terminology for MT. I submit the above definitions for criticism and suggestions.

A Type of Program for Mechanical Translation

J. P. Cleave, University of Southampton, Southampton, England*

A program for the mechanical translation of a limited French vocabulary into English was constructed for operation on the computer APEXC. Its principal features were an improved routine for dictionary look-up, and an organization permitting systematic incorporation of additional subroutines. A program for syntactic processing was constructed but was too large for the available storage space. It examined preceding and following items — stems or endings — in order to choose correct equivalents, and used a dictionary of syntactic sequences or structures to effect local word-order change.

APEXC

The computer has a magnetic drum store with 1024 locations arranged in 32 tracks each of 32 locations. Each location contains 32 bits. Any location can therefore be specified by an address of 10 bits. Both data and instructions are stored on the drum.

An instruction consists of 32 binary digits and specifies an operation (function), the 10 bit address of an operand contained in the store and the address (10 bits) of the next instruction, which again is contained in one location in the store. The arrangement of the digits of an instruction is shown below (Fig. 1).

APEXC has one branch (jump) instruction discriminating between positive (or zero) and negative.

The following abbreviations will be used:

O_x	operand address (X-address) of an instruction O.
O_y	next instruction address (Y-address) of O.
$(O_x)_{ls}$	least significant digit of O_x (i.e., digit 10).
$(O_y)_{ms}$	most significant digit of O_y (i.e., digit 11).
(z)	contents of the location whose address is z.

Dictionary Subroutines

The dictionary procedure is best explained by considering a simplified example with a dictionary of 16 positive entries stored in increasing numerical order in locations 1, 2, 3, ... 16. Suppose W is a word, known to be in the dictionary, whose address in the dictionary is required.

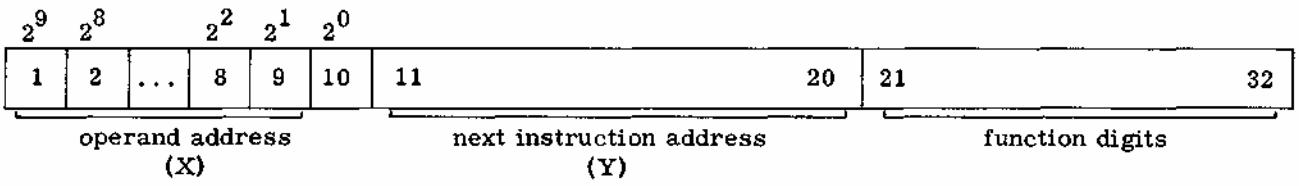


Figure 1

---

* This paper is a report of work done in cooperation with Dr. A. D. Booth and Mr. L. Brandwood at the Computational Laboratory, Birkbeck College, London.

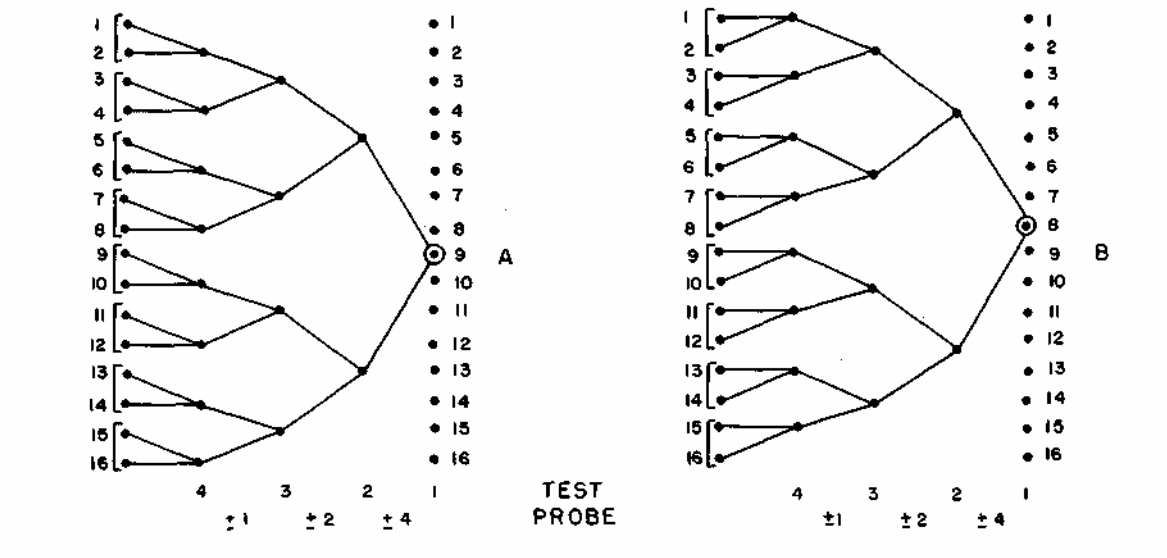


Figure 2

The bracketing procedure[1] requires us to start in the middle of the dictionary, either at 8 or 9. Suppose 8 is chosen; the procedure for 9 is analogous (see Fig. 2).

An "operation" consists of forming W-(y) by means of a subtraction instruction O. If the result is positive, a "probe-number" p is added to O_x, if negative it is subtracted, p is then divided by 2.

The first operation is on (8) (i.e., $O_x = 8$) with $p = 2^2$. After the operation $O_x = 12$ or 4 (i.e., $O_x = 8 + 2^2$ or $8 - 2^2$), the new probe-number is $p = 2^1$.

The second operation gives a new probe-number of 2^0. The third test, therefore, shows W to be in one of the 8 sets of 2 shown in the diagram.

The fourth operation is slightly different from those preceding. It can be seen that operations 1, 2, 3 each discriminate between two new addresses: the fourth discriminates between one new address and one that has been tested before.

If we now examine the dictionary entry specified by O_x at the beginning of operation 4, it can be seen that W is either in O_x or $O_x + 1$. (If the initial location had been 9, the alternatives would be O_x and $O_x - 1$.) Hitherto, dictionary subroutines we have used counted the number of operations performed and at the final operation tested O_x and its neighbor for identity with W. This latter test had to be synthesized and so required several instructions. This disadvantage can be eliminated if the final operation is similar to its predecessors.

Suppose operation 4 is similar to 1, 2, 3. At the conclusion of the third test $p = 2^{-1} = 1/2$. This is a '1' in $(O_y)_{ms}$. The X-addresses formed are shown in Fig. 3.

If the initial location is 9 and $(O_y)_{ms}$ prior to operation 3 is '0', the correct address of W in the dictionary will be formed in O_x. But O_y.. is the address of the next instruction to O in the dictionary routine and is altered by the addition of 2^{-1} to O_x to $O_v' = O_v + 2^9$, thus enabling a jump to occur at precisely the right moment in the sequence of operations. O_y' is the address of the first instruction of the routine following dictionary look-up. If the initial

1. Booth, A. D., "Use of a Computing Machine as a Mechanical Dictionary", Nature, vol. 176, Sept. 17th, 1955, p.565.

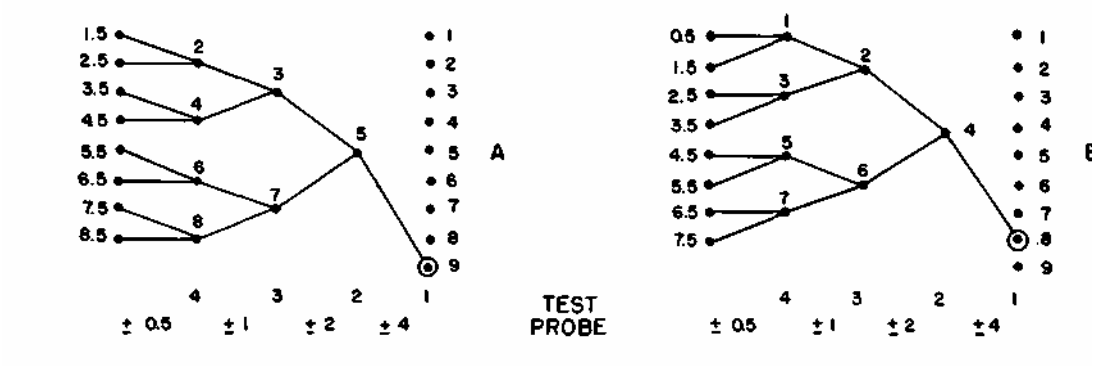


Figure 3

location is 8, W is located correctly only if $(O_y)ms = 1$ Here $O_y' = O_y - 29$

The efficacy of this method clearly depends upon the fact that $(O_x)_{ls}$ is next to $(Oy)_{ms}$ (see Fig. 1). This convenient arrangement now enables us to dispense with special arrangements for the final operation, counting the number of operations performed and special orders for jumping to the next sequence. The dictionary program now occupies only 11 locations: it was used in the MT program explained below.

If the W is not in the dictionary, then this method of dictionary look-up will select the greatest entry less than W.

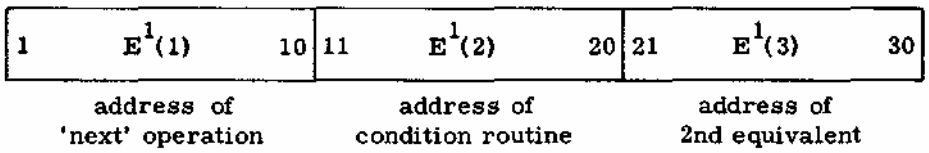
It might be supposed that a further increase of speed could be obtained if during each of the above operations a test for zero is made (i.e., identity between W and the dictionary entry). Suppose a dictionary of 2^n entries. One dictionary entry can be located during the 1st test, 2 during the 2nd, 4 during the 3rd, ... 2^{r-1} during the r^{th} , . . .; $2^{n-1}+1$ requires n tests. (The extra 1 is an entry that cannot be located by a zero test: in the examples of Fig. 2, either 1, or 16.) Assuming that each entry is equally likely to occur in a text, the average number of operations to locate a single word is

$$m = [1.1 + 2.2 + 4.3 + \ldots + r2^{r-1} + \ldots + (n2^{n-1} + n)] /2^n$$
$$= n - 1 + (1 + n)/2^n.$$

Thus if n is large only one operation is saved; the extra programming required in a test for zero is therefore not worth-while with a computer without this facility.

The Basic MT Program

All data to be "recognized" were, with a few exceptions, included in the main dictionary. The input routine compared sequences of symbols between "space" marks with the dictionary entries. This routine therefore had only to recognize a "space" symbol on the input tape. All punctuation marks, and the symbol for the end of text, were included as dictionary entries. Each dictionary entry D of the main- and ending-dictionaries was confined to one storage location and had two equivalents. The second of these, E^2, was the target language equivalent of the dictionary entry. In general E^2 occupied several locations. All "syntactical" operations were performed on the "first equivalents, " E^1 , each of which occupied only one storage location. Each E^1 was constructed uniformly and consisted of three sets of ten digits specifying addresses $E^1(1)$, $E^1(2)$, $E^1(3)$. (See Fig. 4.)



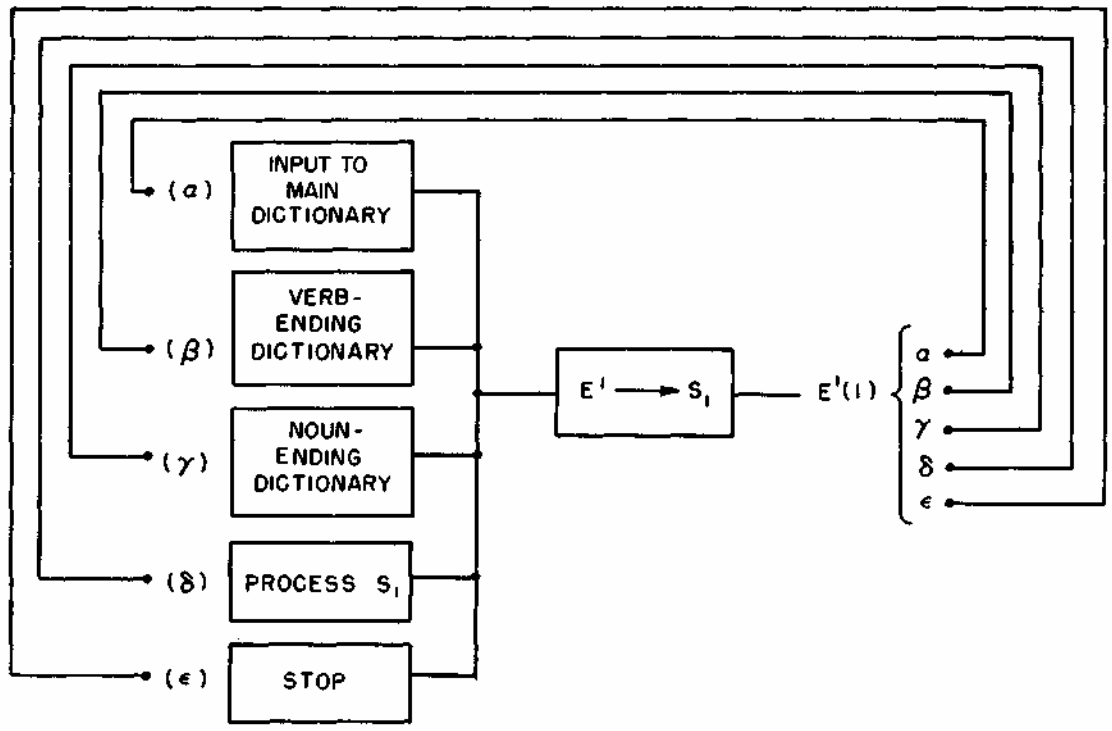
Structure of First Equivalent

Figure 4

The address $E^1(1)$ of E^1 was used in the following manner. After an input datum had been identified (either a complete word, or stem) its first equivalent E^1 was placed in a set S_1 of consecutive storage locations. $E^1(1)$ then specified the address of the first instruction of the next sequence of operations. Thus if a verb were compared with the main dictionary, E_i^1 of the stem was extracted and placed in location n of S_1. $E_i^1(1)$ then determined the next operation. $E_i^1(1)$ for a verb stem was β, the address of the first order of the routine directing the dictionary look-up procedure to the verb-ending dictionary. E_j^1 of the verb ending was then stored in location n + 1 of S_1. $E_j^1(1)$ was the address α of the input routine. Thus after the first equivalent of the verb stem had been stored in S_1, a new word was fed in to be compared with the entries of the main dictionary. The first equivalent of a full stop had address $E^1(1) = S$, the address of the initial instruction of a routine for processing the accumulated data in S. (Fig. 5.) $E^1(1)$ for an end-of-text symbol was ε, a stop order.

A program for processing the first equivalents was constructed but was found to be too large for the available storage space and was abandoned. The plan of this routine, however, will be stated.

The processing of S_1 consisted of carrying out in turn the operations whose first instructions were determined by the second address $E^1(2)$ of each first equivalent in S_1. These operations — condition routines — had two functions. The first was to examine, where necessary, equivalents preceding and following to determine whether $E^1(3)$ specified the correct second equivalent. The second function was to place a code number C corresponding to E in another series of locations S_2. Convenient sub-sequences of the code numbers in S_2 were then compared to a "structure-dictionary." Recognition of these sub-sequences resulted in a rearrangement of the order of the recognized



The Function of First Equivalents

Figure 5

C-sequence and the corresponding E^1 -sequence. The code-numbers were therefore assigned in such a manner that the sequences requiring rearrangement could be recognized distinctly. Although in most cases this assignment coincided with the usual classification of verb, pronoun, etc., there were some C which did not correspond to these categories. Thus donn was entered in the main dictionary, with 'give' as the target language equivalent. The condition routine for this entry assigned a code number ($verb_1$) to it. erons was an entry in the verb-ending dictionary. The condition routine determined by its first equivalent gave it a code number ($verb_2$). The second equivalent of erons was 'will'. Thus when donnerons occurred in the input text, the first equivalents of donn and erons were placed in consecutive locations in S_1. When the condition routines were operated, the code numbers ($verb_1$) and ($verb_2$) were placed in order in S_2. Following these routines the structure dictionary recognized the sequence ($verb_1$) ($verb_2$) as one requiring transposition. The corresponding data in S_1 were then transposed. Thus the final printing operation printed the target language equivalents of donn/erons in reverse order to yield 'will give'. This procedure was used to perform the pronoun-verb inversion.

The final stage of the program was a routine for printing the second equivalents. In the program which was put on APEXC the processing of S_1 was omitted so that the dictionary routines were immediately followed by the print routine. The print routine printed the contents of the addresses specified by the 3rd address of the first equivalents in S_1. Each location containing a second equivalent also contained an indication of whether the content of the next location was also to be printed. By this means equivalents of any desired length could be printed.

Some Characteristics of the Program

This program had two important features.

Firstly, all operations within the program were carried out on the first equivalents. As these were uniformly constructed, a greater simplicity was achieved than if the foreign language words or target language words had been processed directly.

Secondly, the distinct parts of the whole program were isolated, the linkages being supplied by the addresses in the first equivalents. Thus extra subroutines could be constructed and linked to the program merely by altering addresses in the relevant first equivalents. For instance, if a more refined condition routine was necessary for a certain set of first equivalents, this routine could be placed in the store and the second addresses of the first equivalents altered to the address of the initial order of the new routine.

The size of storage in the computer imposed severe limits on the extent and performance of the program. Thus very small dictionaries were used, although best use was made of the space available by means of stem-ending splitting. Apart from these faults, there were two inherent drawbacks of the above type of program.

The use of separate condition routines employing a matching procedure to examine the minor context of a first equivalent lead to an excessive program. A more economical approach would be to calculate correct alternatives from code numbers by some means. This would greatly reduce the storage space assigned to this particular part of the program.

Secondly, the method of effecting change of word order appears to be applicable only to subsections of languages where permutation of target language order into foreign language order is purely local. Thus if a set of n consecutive code numbers in S_2 was matched by the above method to a dictionary of structures, the change of word order was confined to the corresponding set of n first equivalents only. This process was clearly incapable of dealing directly with rearrangements of blocks of words.

A possible solution of the problem here would be to use two structure-dictionaries, one for permuting elements within a block, another to permute the blocks. The necessity of using a structure-dictionary will disappear when a suitable technique of calculation (as opposed to matching) has been discovered.

A Framework for Syntactic Translation †

V. H. Yngve, Massachusetts Institute of Technology, Cambridge, Massachusetts

Adequate mechanical translation can be based only on adequate structural descriptions of the languages involved and on an adequate statement of equivalences. Translation is conceived of as a three-step process: recognition of the structure of the incoming text in terms of a structural specifier; transfer of this specifier into a structural specifier in the other language; and construction to order of the output text specified.

Introduction

THE CURRENT M.I.T. approach to mechanical translation is aimed at providing routines intrinsically capable of producing correct and accurate translation. We are attempting to go beyond simple word-for-word translation; beyond translation using empirical, ad hoc, or pragmatic syntactic routines. The concept of full syntactic translation has emerged: translation based on a thorough understanding of linguistic structures, their equivalences, and meanings.

The Problems

The difficulties associated with word-for-word translation were appreciated from the very beginning, at least in outline form. Warren Weaver[1] and Erwin Reifler[2] in early memoranda called attention to the problems of multiple meaning, while Oswald and Fletcher[3] began by fixing their attention on the word-order problems — particularly glaring in the case of German-to-English word-for-word translations. Over the years it has become increasingly clear that most, if not all, of the problems associated with word-for-word translation can be solved by the proper manipulation or utilization of the context. Context is to be understood here in its broadest interpretation. Contextual clues were treated in detail in an earlier article.[4] The six types of clues discussed there will be reformulated briefly here. They are:

1) The field of discourse. This was one of the earliest types of clues to be recognized. It can, by the use of specialized dictionaries, assist in the selection of the proper meaning of words that carry different meanings in different fields of discourse. The field of discourse may be determined by the operator, who places the appropriate glossary in the machine; or it may be determined by a machine routine on the basis of the occurrences of certain text words that are diagnostic of the field.

---

† This work was supported in part by the U. S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research); and in part by the National Science Foundation.

1. Warren Weaver, "Translation," Machine Translation of Languages, edited by Locke and Booth (New York and London, 1955)

2. Erwin Reifler, "Studies in Mechanical Translation No. 1, MT, " mimeographed (Jan. 1950)

3. Oswald and Fletcher, "Proposals for the Mechanical Resolution of German Syntax Patterns, " Modern Language Forum, vol. XXXVI, no. 2-4 (1951)

4. V. H. Yngve, "Terminology in the Light of Research on Mechanical Translation, " Babel, vol. 2, no. 3 (Oct. 1956)

2) Recognition of coherent word groups, such as idioms and compound nouns. This clue can provide a basis for translating such word groups correctly even when their meaning does not follow simply from the meanings of the separate words.

3) The syntactic function of each word. If the translating program can determine syntactic function, clues will be available for solving word-order problems as well as a large number of difficult multiple-meaning problems. Clues of this type will help, for example, in determining whether der in German should be translated as an article or as a relative or demonstrative pronoun, and whether it is nominative, genitive, or dative. They will also assist in handling the very difficult problems of translating prepositions correctly.

4) The selectional relations between words in open classes, i.e., nouns, verbs, adjectives, and adverbs. These relations can be utilized by assigning the words to various meaning categories in such a way that when two or more of these words occur in certain syntactic relationships in the text, the correct meanings can be selected.

5) Antecedents. The ability of the translating program to determine antecedents will not only make possible the correct translation of pronouns, but will also materially assist in the translation of nouns and other words that refer to things previously mentioned.

6) All other contextual clues, especially those concerned with an exact knowledge of the subject under discussion. These will undoubtedly remain the last to be mechanized.

Finding out how to use these clues to provide correct and accurate translations by machine presents perhaps the most formidable task that language scholars have ever faced.

Two Approaches

Attempts to learn how to utilize the above-mentioned clues have followed two separate approaches. One will be called the "95 per cent approach" because it attempts to find a number of relatively simple rules of thumb, each of which will translate a word or class of words correctly about 95 per cent of the time, even though these rules are not based on a complete understanding of the problem. This approach is used by those who are seeking a short-cut to useful, if not completely adequate, translations. The other approach concentrates on trying to obtain a complete understanding of each portion of the problem so that completely adequate routines can be developed.

At any stage in the development of mechanical translation there will be some things that are perfectly understood and can therefore serve as the basis for perfect translation. In the area of verb, noun, and adjective inflection, it is possible to do a "100 per cent job" because all the paradigms are available and all of the exceptions are known and have been listed. In this area one need not be satisfied with anything less than a perfect job.

At the same time there will be some things about language and translation that are not understood. It is in this area that the difference between the two approaches shows up. The question of when to translate the various German, French, or Russian verb categories into the different sets of English verb categories is imperfectly understood. Those who adopt the 95 per cent approach will seek simple partial solutions that are right a substantial portion of the time. They gain the opportunity of showing early test results on a computer. Those who adopt the 100 per cent approach realize that in the end satisfactory mechanical translation can follow only from the systematic enlarging of the area in which we have essentially perfect understanding.

The M.I. T. group has traditionally concentrated on moving segments of the problem out of the area where only the 95 per cent approach is possible into the area where a 100 per cent approach can be used. Looking at mechanical translation in this light poses the greater intellectual challenge, and we believe that it is here that the most significant advances can be made.

Syntactic Translation

Examination of the six types of clues mentioned above reveals that they are predominantly concerned with the relationships of one word to another in patterns. The third type — the ability of the program to determine the syntactic function of each word — is basic to the others. It is basic to the first: If the machine is to determine correctly the field of discourse at every point in the text, even when the field changes within one sentence, it must use the relationship of the words in syntactic patterns as the key for finding which words refer to which field. It is basic to the second because idioms, noun compounds, and so on, are merely special patterns of words that stand out from

more regular patterns. It is basic to the fourth because here we are dealing with selectional relationships between words that are syntactically related. It is basic to the fifth because the relationship of a word to its antecedent is essentially a syntactic relationship. It is probably even basic to the last, the category of all other contextual clues.

Any approach to mechanical translation that attempts to go beyond mere word-for-word translation can with some justification be called a syntactic approach. The word "syntactic" can be used, however, to cover a number of different approaches. Following an early suggestion by Warren Weaver,[1] some of these take into consideration only the two or three immediately preceding and following words. Some of them, following a suggestion by Bar-Hillel,[5] do consider larger context, but by a complicated scanning forth and back in the sentence, looking for particular words or particular diacritics that have been attached to words in the first dictionary look-up. To the extent that these approaches operate without an accurate knowledge and use of the syntactic patterns of the languages, they are following the 95 per cent approach.

Oswald and Fletcher[3] saw clearly that a solution to the word-order problems in German-to-English translation required the identification of syntactic units in the sentence, such as nominal blocks and verbal blocks. Recently, Brandwood[6] has extended and elaborated the rules of Oswald and Fletcher. Reifler,[7] too, has placed emphasis on form classes and the relationship of words one with the other. These last three attempts seem to come closer to the 100 per cent way of looking at things.

Bar-Hillel,[8] at M.I.T., introduced a 100 per cent approach years ago when he attempted to adapt to mechanical translation certain ideas of the Polish logician Ajdukiewicz. The algebraic notation adopted for syntactic categories, however, was not elaborate enough to express the relations of natural languages.

Later, the author [9, 10] proposed a syntactic method for solving multiple-meaning and word-order problems. This routine analyzed and translated the input sentences in terms of successively included clauses, phrases, and so forth.

More recently, Moloshnaya [11] has done some excellent work on English syntax, and Zarechnak[12] and Pyne[13] have been exploring with Russian a suggestion by Harris[14] that the text be broken down by transformations into kernel sentences which would be separately translated and then transformed back into full sentences. Lehmann,[15] too, has recently emphasized that translation of the German noun phrase into English will require a full descriptive analysis.

---

5. Y. Bar-Hillel, "The Present State of Research on Mechanical Translation, " American Documentation, 2:229-237 (1951)

6. A. D. Booth, L. Brandwood, J. P. Cleave, Mechanical Resolution of Linguistic Problems, Academic Press (New York, 1958)

7. Erwin Reifler, "The Mechanical Determination of Meaning, " Machine Translation of Languages, edited by Locke and Booth (New York and London, 1955)

8. Y. Bar-Hillel, "A Quasi-Arithmetical Notation for Syntactic Description, " Language, vol. 29, no. 1 (1953)

9. V. H. Yngve, "Syntax and the Problem of Multiple Meaning," Machine Translation of Languages, edited by Locke and Booth (New York and London, 1955)

10. V. H. Yngve, "The Technical Feasibility of Translating Languages by Machine," Electrical Engineering, vol. 75, no. 11 (1956)

11. T. N. Moloshnaya, "Certain Questions of Syntax in Connection with Machine Translation from English to Russian," Voprosy Yazykoznaniya. no. 4 (1957)

12. M. M. Zarechnak, "Types of Russian Sentences," Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies, Georgetown University (1957)

13. J. A. Pyne, "Some Ideas on Inter-structural Syntax," Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies, Georgetown University (1957)

14. Z. S. Harris, "Transfer Grammar," International Journal of American Linguistics, vol. XX, no. 4 (Oct. 1954)

15. W. P. Lehmann, "Structure of Noun Phrases in German," Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies, Georgetown University (1957)

In much of the work there has been an explicit or implicit restriction to syntactic relationships that are contained entirely within a clause or sentence, although it is usually recognized that structural features, to a significant extent, cross sentence boundaries. In what follows, we will speak of the sentence without implying this restriction.

The Framework

The framework within which we are working is presented in schematic form in Fig. 1. This framework has evolved after careful consideration of a number of factors. Foremost among these is the necessity of breaking down a problem as complex as that of mechanical translation into a number of problems each of which is small enough to be handled by one person.

Figure 1 represents a hypothetical translating machine. German sentences are fed in at the left. The recognition routine, R.R., by referring to the grammar of German, G_1, analizes the German sentence and determines its structural description or specifier, S_1, which contains all of the information that is in the input sentence. The part of the information that is implicit in the sentence (tense, voice, and so forth) is made explicit in S_1. Since a German sentence and its English translation generally do not have identical structural descriptions, we need a statement of the equivalences, E, between English and German structures, and a structure transfer routine, T.R., which consults E and transfers S_1 into S_2, the structural description, or specifier, of the English sentence. The construction routine, C.R., is the routine that takes S_2 and constructs the appropriate English sentence in conformity with the grammar of English, G_2.

This framework is similar to the one previously published[16] except that now we have added the center boxes and have a much better understanding of what was called the "message" or transition language — here, the specifiers. Andreyev[17] has also recently pointed out that translation is essentially a three-step process and that current published proposals have combined the first two steps into one. One might add that some of the published proposals even try to combine all three steps into one. The question of whether there are more than three steps will be taken up later.

A few simple considerations will make clear why it is necessary to describe the structure of each language separately. First, consider the regularities and irregularities of declensions and conjugations. These are, of course, entirely relative to one language.

Context, too, is by nature contained entirely within the framework of one language. In considering the translation of a certain German verb form into English, it is necessary to understand the German verb form as part of a complex of features of German structure including possibly other verb forms within the clause, certain adverbs, the structure of neighboring clauses, and the like. In translating into English, the appropriate complex of features relative to English structure must be provided so that each verb form is understood correctly as a part of that English complex.

The form of an English pronoun depends on its English antecedent, while the form of a German pronoun depends on its German antecedent — not always the same word because of the multiple-meaning situation. As important as it is to locate the antecedent of the input pronoun in the input text, it is equally important to embed the output pronoun in a proper context in the output language so that its antecedent is clear to the reader.

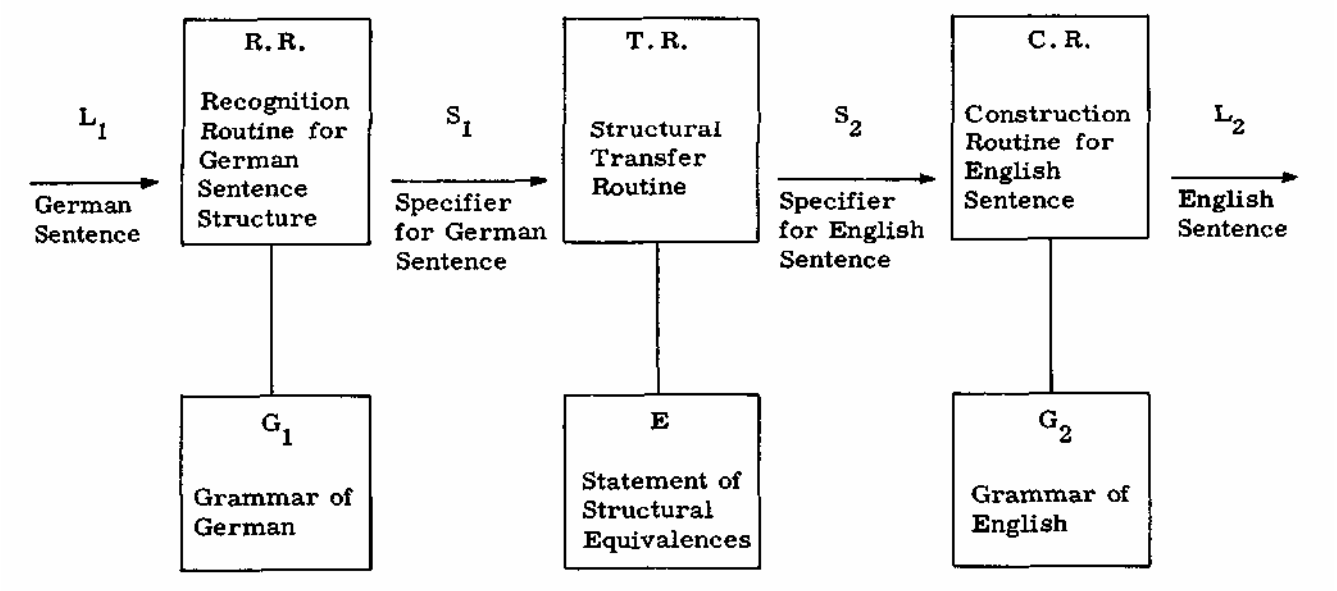
In all of these examples it is necessary to understand the complete system in order to program a machine to recognize the complex of features and to translate as well as a human translator. If one is not able to fathom the complete system, one has to fall back on hit-or-miss alternative methods — the 95 per cent approach. In order to achieve the advantages of full syntactic translation, we will have to do much more very careful and detailed linguistic investigation.

Stored Knowledge

The diagram (Fig. 1) makes a distinction between the stored knowledge (the lower boxes) and the routines (the upper boxes). This distinction represents a point of view which may be academic: In an actual translating program the routine boxes and the stored knowledge boxes might be indistinguishable. For our purpose, however, the lower boxes represent our

16. V. H. Yngve, "Sentence-for-sentence Translation," MT, vol. 2, no. 2 (1955)

17. N. D. Andreyev, "Machine Translation and the Problem of an Intermediary Language, Voprosy Yazykoznaniya, no. 5 (1957)



A Framework for Mechanical Translation

Figure 1

knowledge of the language and are intended not to include any details of the programming or, more particularly, any details of how the information about the languages is used by the machine. In other words, these boxes represent in an abstract fashion our understanding of the structures of the languages and of the translation equivalences. In an actual translating machine, the contents of these boxes will have to be expressed in some appropriate manner, and this might very well take the form of a program written in a pseudo code, programmable on a general-purpose computer. Earlier estimates[9] that the amount of storage necessary for syntactic information may be of the same order of magnitude as the amount of storage required for a dictionary have not been revised.

Construction

The Construction Routine, C.R. in Figure 1, constructs to order an English sentence on the prescription of the specifier, S_2. It does this by consulting its pharmacopoeia, the grammar of English, G_2, which tells it how to mix the ingredients to obtain a correct and grammatical English sentence, the one prescribed.

The construction routine is a computer program that operates as a code conversion device, converting the code for the sentence, the specifier, into the English spelling of the sentence . The grammar may be looked upon in this light as a code book, or, more properly, as an algorithm for code conversion. Alternately the construction routine can be regarded as a function generator. The independent variable is the specifier, and the calculated function is the output sentence. Under these circumstances, the grammar, G_2, represents our knowledge of how to calculate the function.

The sentence construction routine resembles to some extent the very suggestive sentence generation concept of Chomsky,[18] but there is an important difference. Where sentence generation is concerned with a compact representation of the sentences of a language, sentence construction is concerned with constructing, to order, specified sentences one at a time. This difference in purpose necessitates far-reaching differences in the form of the grammars.

---

18. Noam Chomsky, Syntactic Structures, Mouton and Co., 'S-Gravenhage (1957)

Specifiers

For an input to the sentence construction routine, we postulated an encoding of the information in the form of what we called a specifier. The specifier of a sentence represents that sentence as a series of choices within the limited range of choices prescribed by the grammar of the language. These choices are in the nature of values for the natural coordinates of the sentence in that language. For example: to specify an English sentence, one may have to specify for the finite verb 1st, 2nd, or 3rd person, singular or plural, present or past, whether the sentence is negative or affirmative, whether the subject is modified by a relative clause, and which one, etc. The specifier also specifies the class to which the verb belongs, and ultimately, which verb of that class is to be used, and so on, through all of the details that are necessary to direct the construction routine to construct the particular sentence that satisfies the specifications laid down by the author of the original input sentence.

The natural coordinates of a language are not given to us a priori, they have to be discovered by linguistic research.

Ambiguity within a language can be looked at as unspecified coordinates. A writer generally can be as unambiguous as he pleases — or as ambiguous. He can be less ambiguous merely by expanding on his thoughts, thus specifying the values of more coordinates. But there is a natural limit to how ambiguous he can be without circumlocutions. Ambiguity is a property of the particular language he is using in the sense that in each language certain types of ambiguity are not allowed in certain situations. In Chinese, one can be ambiguous about the tense of verbs, but in English this is not allowed: one must regularly specify present or past for verbs. On the other hand, one is usually ambiguous about the tense of adjectives in English, but in Japanese this is not allowed.

It may be worth while to distinguish between structural coordinates in the narrow sense and structural coordinates in a broader, perhaps extra linguistic sense, that is, coordinates which might be called logical or meaning coordinates. As examples, one can cite certain English verb categories: In a narrow sense, the auxiliary verb 'can' has two forms, present and past. This verb, however, cannot be made future or perfect as most other verbs can. One does not say 'He has can come,' but says, instead, 'He has been able to come,' which is structurally very different. It is a form of the verb 'to be' followed by an adjective which takes the infinitive with 'to.' Again the auxiliary 'must' has no past tense and again one uses a circumlocution — 'had to.' If we want to indicate the connection in meaning (paralleling a similarity in distribution) between 'can' and 'is able to' and between 'must' and 'has to,' we have to use coordinates that are not structural in the narrow sense. As another example, there is the use of the present tense in English for past time (in narratives), for future time ('He is coming soon'), and with other meanings. Other examples, some bordering on stylistics, can also be cited to help establish the existence of at least two kinds of sentence coordinates in a language, necessitating at least two types of specifiers.

A translation routine that takes into consideration two types of specifiers for each language would constitute a five-step translation procedure. The incoming sentence would be analyzed in terms of a narrow structural specifier. This specifier would be converted into a more convenient and perhaps more meaningful broad specifier, which would then be converted into a broad specifier in the other language, then would follow the steps of conversion to a narrow specifier and to an output sentence.

Recognition

One needs to know what there is to be recognized before one can recognize it. Many people, including the author, have worked on recognition routines. Unfortunately, none of the work has been done with the necessary full and explicit knowledge of the linguistic structures and of the natural coordinates.

The question of how we understand a sentence is a valid one for linguists, and it may have an answer different from the answer to the question of how we produce a sentence. But it appears that the description of a language is more easily couched in terms of synthesis of sentences than in terms of analysis of sentences. The reason is clear. A description in terms of synthesis is straightforward and unambiguous. It is a one-to-one mapping of specifiers into sentences. But a description in terms of analysis runs into all of the ambiguities of language that are caused by the chance overlapping of different patterns: a given sentence may be understandable in terms of two or more different specifiers. Descriptions in terms of analysis will probably not be available until after we

have the more easily obtained descriptions in terms of synthesis.

The details of the recognition routine will depend on the details of the structural description of the input language. Once this is available, the recognition routine itself should be quite straightforward. The method suggested earlier by the author[9] required that words be classified into word classes, phrases into phrase classes, and so on, on the basis of an adequate descriptive analysis. It operated by looking up word-class sequences, phrase-class sequences, etc., in a dictionary of allowed sequences.

Transfer of Structure

Different languages have different sets of natural coordinates. Thus the center boxes (Fig. 1) are needed to convert the specifiers for the sentences of the input language into the specifiers for the equivalent sentences in the output language. The real compromises in translation reside in these center boxes. It is here that the difficult and perhaps often impossible matching of sentences in different languages is undertaken. But the problems associated with the center box are not peculiar to mechanical translation. Human translators also face the very same problems when they attempt to translate. The only difference is that at present the human translators are able to cope satisfactorily with the problem.

We have presented a framework within which work can proceed that will eventually culminate in mechanical routines for full syntactic translation. There are many aspects of the problem that are not yet understood and many details remain to be worked out. We need detailed information concerning the natural coordinates of the languages. In order to transfer German specifiers into English specifiers, we must know something about these specifiers. Some very interesting comparative linguistic problems will undoubtedly turn up in this area.

The author wishes to express his indebtedness to his colleagues G. H. Matthews, Joseph Applegate, and Noam Chomsky, for some of the ideas expressed in this paper.

Order of Subject and Predicate in Scientific Russian†

Ilse Lehiste, University of Michigan, Ann Arbor, Michigan

A study by Kenneth E. Harper indicates that word order in Russian scientific writing is sufficiently similar to that of English to permit word-for-word translation from Russian to English. Further study of Russian texts shows that word order in scientific Russian is sufficiently different to require analysis, for translation purposes, based on form and function rather than on word-for-word correspondence.

IN HIS "A Preliminary Study of Russian",[1] Kenneth E. Harper states that a "word-for-word translation of Russian is adequate for understanding," since "in the field of scientific writing, Russian sentence structure is definitely close to English — much closer than is normal for other forms of Russian prose. "

In support of this statement, Harper quotes certain figures:

"From a sample of 1, 528 sentences containing a subject and verb:

Subject before verb: 81% of all occurrences
Verb before subject: 19% of all occurrences

(195 additional sentences contained an impersonal, or understood, subject; 24 sentences contained no verb.) The position of subject before verb (normal English word order) thus appears to prevail approximately four-fifths of the time."

Proceeding from these assumptions, Harper builds his system of mechanical translation of Russian upon word-for-word translation, stripping the Russian words of their endings to identify them by their stems, which are listed in the dictionary.

The purpose of this paper is to verify to what extent these assumptions are valid, i.e. to determine in what measure word order is predictable in scientific Russian.

One hundred twenty-eight pages of continuous text[2] were analyzed for the relative positions of the subject and the predicate. The predicate spot was determined syntactically, by its function, and the following types of fillers were found in the predicate spot: verb, adjective, noun, prepositional phrase, and various types of impersonal expressions.[3] Sentences containing no predicate (so-called "nominal sentences") were not analyzed; their number was found to be relatively insignificant (headings, titles, bibliography lists, etc.). Main clauses and dependent clauses were not separated in the analysis.

Out of a total of 2914 clauses thus analyzed, the word order was as follows:

Subject — Predicate in 1915 instances, or 65.71% of the total;
Predicate — Subject in 342 instances, or 11.74% of the total.

---

† This study was conducted at the University of Michigan with research funds provided by the Engineering Research Institute.

1. Machine Translation of Languages, edited by W. N. Locke and A. D. Booth, John Wiley and Sons, Inc., New York, 1955, pp.66-85.

2. Zhurnal eksperimental'noy i teoreticheskoy fiziki, Tom 28, 1955, vyip. 1.

3. The classification is based on the Grammatika russkogo jazyka of the Academy of Sciences of the U.S. S.R., Moscow, 1954, Vol. II, 1, p.387ff.

The clause contained no subject in 657 instances, or 22.55% of the total.

1. The predicate slot was filled by a verb in 1527 instances, or 52. 40% of the total. Of these the word order was Subject — Predicate in 1282 instances, 43.99% of the total; the word order was Predicate — Subject in 245 instances, 8.41% of the total, the ratio being 1282/245, or approximately 5/1.

2. The predicate slot was filled by a noun in 232 instances, or 7.96% of the total. The word order was Subject - Predicate in all instances without exception.

3. The predicate slot was filled by an adjective in 496 instances, or 17. 02% of the total. Of these, the word order was Subject — Predicate in 399 instances, 13.69% of the total; the word order was Predicate — Subject in 97 instances, 3. 33% of the total, the ratio being 399/97, or approximately 4/1.

The adjective filler was subdivided into adjective proper and past participle. The data are as follows:

Predicate slot filled by adjective proper;
Subject - Predicate, 267 instances or 9.16% of the total;
Predicate — Subject, 25 instances or 0. 86% of the total.

Ratio 267/25, or approximately 10/1.
The total number of instances when the predicate slot was filled by adjective proper was 292, or 10.02% of the total.

4. The predicate slot filled by past participle:
Subject — Predicate, 132 instances or 4.53% of the total;
Predicate — Subject, 72 instances or 2.47% of the total.

The ratio was 132/72, or approximately 2/1. The total number of instances when the predicate slot was filled by past participle was 204, or 7.00% of the total.

5. The clauses contained no subject in 657 instances, or 22.55% of the total. Of that number, the predicate slot was filled by an impersonal expression (such as можно, следует, необходимо) in 383 instances, or 13.14%; the predicate slot was filled by a verb with included subject (such as получаем, выражаю) in 226 instances, or 7.76%.

6. The clause contained no other predicative element except an infinitive (strictly speaking, infinitive phrases, introduced by если or чтобы) in 48 instances, or 1.65% of the total.

7. The predicate slot was filled by a prepositional phrase in 2 instances, or 0.07% of the total.

These figures differ considerably from those obtained by Harper. Only approximately 50% of the sentences contain both a subject and a verb. The so-called "normal English word order" occurs in only approximately 44% of actual sentences, as compared to the 81% suggested by Harper. The predicate spot can be filled by a variety of classes of words. Almost 1/4 of the clauses contain no subject. The results of the above study indicate that the word order in scientific Russian is sufficiently different from that of English to make it imperative that the analysis be based on a consideration of form and function rather than word-for-word correspondence.

Semantic Ambiguity

Kenneth E. Harper, University of California, Los Angeles, California

The extent of the problem of multiple meaning in translation is illustrated in this analysis of a sample page of Russian scientific text. The use of an idio-glossary represents only a partial solution to the problem.

AN ANALYSIS of a sample page of Russian scientific text[1] revealed the following distribution of words, with respect to semantic clarity or ambiguity:

	Single Value		Idioms	Multiple Values
	Word-for-word MT	Clarified by structural analysis		
Words	77*	7	24	46 (30% of total)
Running words		40	24	75 (28% of total)

Total number of words: 151
Total number of running words: 266

Figure 1

Callaham's Russian-English Technical and Scientific Dictionary was consulted for English equivalents of the multiple-valued words. The average number of equivalents for these words was 8.6. Many of these equivalents may be considered as synonyms; when there is a fairly distinct change in meaning, the listing is divided into groups, separated by a semi-colon. The average number of such groups, representing distinct meanings, is 3. 0 (see Fig. II).

Conclusion: 30% of the total words in the passage analyzed should be represented by three English equivalents. This figure has no meaning as applied to any given word; it is perhaps an indication of the extent of the problem of multiple meaning. The problem can be partially solved by an arbitrary selection of a given equivalent for certain fields (the Idio-glossary). Even here, there are definite limits, as the list of typical multiple meaning words in Fig. II shows.

---

1. F. M. Gol'tsman and SH. SH. Raskin, "O dielectricheskikh svoistvakh nekotorykh polimorfnykh organicheskikh soedinenii", Doklady Akademii Nauk SSSR, vol. LXXIX, no. 5. 1953. p. 187.

* Of this number, 22 words (42 runnings words) were "technical" words of one value, such as 'polymorphic', 'dielectric', etc. These words composed 15% of the total.

Multiple listings of equivalents, taken from Callaham's Technical Dictionary, are illustrated by the following sample entries. (The figures indicate, left, the total number of equivalents, and right, the lexicographer's attempt to distinguish between groups of synonyms.)

9-3 изменение:
change, alteration, variation, modification, conversion, transformation; fluctuation, deviation; correction

9-2 объем:
volume, size, bulk, space, capacity, contents; compass, extent, amplitude

11-4 переход:
transition, passing over, passing, conversion; passage, crossing, migration (of ions); exchange (of places), switching; blending, shading (of colors)

15-6 величина:
size, dimension, measure; (math.) value, magnitude, quantity, amount; volume, bulk; degree, extent (of error), scope; intensity (of force, etc.); bigness, greatness

9-3 смешанный
mixed, miscellaneous, composite, compound, combination, blended; hybrid; stirred, agitated

7-4 строение:
building, construction; formation; structure; constitution; texture, grain

5-4 явление:
phenomenon; effect; (med.) symptom; appearance, occurrence

6-3 увеличение:
increase, growth, augmentation; enhancement; enlargement, magnification

5-2 потеря:
loss, disappearance, waste, escape (of gas, etc.); (mil.) casualty

7-2 скачок:
jump, skip, leap, spring, bound; rapid change, drop

8-2 нарушение:
breaking, breach, infringement, infraction, transgression, violation; disturbance, dislocation

6-2 отличаться:
differ, be distinguished (by), be characterized (by); surpass, out-do, excel

Figure 2

Contextual Analysis

Kenneth E. Harper, University of California, Los Angeles, California

Ambiguity, both syntactic and semantic, a problem that arises in the translation of Russian to English because of polysemantic forms in Russian, can be resolved by an analysis of the context in which the polysemantic form occurs. This requires a systematic study of context so that word classes which determine the value of ambiguous forms can be established.

IN THE VARIOUS PROPOSALS for word-for-word machine translation of Russian scientific literature into English, each word in the sentence is considered as a separate entity. If a word has more than one English equivalent, or more than one possible syntactic value, the alternatives must be listed. The chief difficulty with the resulting translation is its prolixity: the reader finds himself confronted with numerous alternatives, both syntactic and semantic, in every sentence. The extent of the problem of ambiguity is suggested by the following figures: from a sample Russian scientific text, 43% of the running words were found to be polysemantic (this in addition to syntactic ambiguities which the reader must solve on the basis of numerous alternatives given him in every sentence).

Context

The difficulty with word-for-word translation, then, is that it is really "words-for-word translation". [1] The solution to the problem lies in the reduction of the number of choices confronting the reader by the mechanical selection of the proper (or actual) syntactical and semantic equivalent from the various potential equivalents. Obviously, the solution can be attempted along lines as infinitely complex as those involved in "human translation", in which judgments are based on "context", experience and even upon "taste". Of these the element of "context" is, to some degree, determinable by mechanical means. In its general sense, context signifies environment, i.e., surrounding words in a sentence, surrounding sentences and paragraphs, extending to the broad category of subject areas. The question arises: Is some more limited use of context analysis possible in MT, and how effective is such analysis in the removal of ambiguity?

In an attempt to answer this question, the potentialities of a "contextual analysis" of each ambiguous word (syntactically or semantically ambiguous) have been studied, such analysis to be limited to immediately contiguous words. Thus, for a given ambiguous word (x), reference may be made to the preceding word (x-1) or to the following word (x+1). (In specified instances, reference may be made to words which are separated by neutral words from (x) word.)

The value of this limited contextual analysis was suggested by the inflectional nature of the Russian language. For example, the English preposition, 'of, indicating possession, does not have a "word equivalent" in Russian; the 'of' is generated by the genitive case of the noun or pronoun (<u>добавление смеси</u> = 'the addition <u>of</u> the mixture'). Two difficulties arise

---

1. The problem of word order is not critical in MT, particularly for technical material. Even in the general literary language, the word order, subject-verb-direct object, is preserved in 85 - 90% of all sentences (according to a study of 5000 pages of Russian prose text, cited in <u>Voprosy grammaticheskogo stroya,</u> Izdatel'stvo Akademii Nauk SSSR, Moscow, 1955, p. 471).

in straight word-for-word MT: 1) difficulty of identifying the genitive ending for most nouns, so that the above Russian words may theoretically mean 'the addition to the mixture', 'the addition the mixture', or 'the addition the mixtures', as well as the translation given above; 2) the 'of generated by the genitive case is often disregarded, under the condition, for example, that the word is preceded by a preposition which governs the genitive case. The task of deciding whether or not to retain the 'of' falls upon the reader. The problem results, of course, from the syntactical compactness of inflected languages. Since syntactical information in Russian is contained not in discrete items (individual inflected words), but in the relationship between words, a comparison process is imperative.

A second reason for believing in the potential of contextual analysis is the effect that consideration of immediately contiguous words has upon the removal of semantic ambiguity of a given word. Professor Kaplan's study on this problem suggests that a marked reduction of ambiguity is the result of considering one or two words preceding and following the ambiguous English word.[2] This is a completely-virgin field of investigation, but preliminary studies indicate that within a closed area of discourse, such as Russian technical literature, the problem of multiple meaning can be satisfactorily handled through the analysis of contiguous words.

In the two following sections studies on the effect of syntactic and semantic clarification by this method are summarized.

Clarification of Syntax

It is essential in this system that any given word in a Russian sentence be subject to retention and further inspection; in other words, location of the item in the memory is only (or nay be only) half the job. Even after its grammatical features have been determined, whether in a paradigm or stem-affix machine dictionary, :he word is not to be printed by the output den.ce until a "go ahead" signal is given. In theory, every word in a sentence is potentially useful to a contiguous word; every word is a potential determiner, and, if it is in any way ambiguous, a potential determinee. Our problem is to discover the manner in which this relationship is expressed, and to represent it in codable form. In certain instances, as in the relationship between adjective and noun, for example, the mutual influence is recognizable in terms of conventional grammar; more frequently, the relationship is unpredictable and must be discovered by observation of behavior in a large number of situations. In any event, the ability to make reference to words in immediate contiguity is inherent to this system.

For purposes of syntactical clarification, conventional grammatical concepts are quite useful. It is helpful, for instance, to have available, in coded form, the following information for words in a Russian sentence: part of speech of all words; case, number, and gender of nouns; the infinitive form and tense of verbs; case and number of certain adjectives, etc. Reference to this information may be helpful in contextual analysis. It should be stressed that reference is made to these coded features, rather than to "the word" itself. In the latter process, we become involved in the identification of idioms, i.e., in the problem of lexical relationship; our present interest is in the structural relationship and its effect upon clarification of syntax.

The processing of syntactically ambiguous words may be summarized in the following descriptive terms:

1) Nouns

a) Genitive Case

For masculine nouns, this case is identifiable by ending (disregarding, in technical Russian, the almost non-existent animate noun). For all neuter and feminine nouns, this case is ambiguous by ending in the singular. For all unmodified nouns which are definitely or potentially genitive case, by ending, the English preposition 'of' is generated only under the condition that the preceding word is a noun. The 'of' is to precede the noun identified as genitive; if adjectives precede the noun in question, the 'of' is to precede all such modifiers. In referring to the part of speech of the preceding word, modifiers of the word in question are ignored.

добавление смеси
= 'addition (of) the mixture'

добавление этой смеси
= 'addition (of) this mixture'

2. Kaplan, Abraham, "An Experimental Study of Ambiguity and Context", Mechanical Translation, vol. 2, no. 2, pp. 39-46, November 1956.

The result of the above restriction (that the preceding word must be a noun) automatically eliminates the generation of the 'of' in the frequent instances where the genitive case is required by Russian grammatical rules, but where its identification only serves to hinder the translation, — for example, when the preceding word is: a preposition, a cardinal number, a comparative adjective, a negative (нет), a verb which governs the genitive case, words of quantity (много, сколько), negated verb, etc.

This rule, formulated purely on the basis of observed behavior, very accurately approximates the control over "context" unconsciously enjoyed by the human reader of Russian.

b) Instrumental Case

This case is not ambiguous by ending. Nouns in this case (and any preceding modifiers) are to be preceded by the English word 'by' ('with' in certain specified cases), except when the preceding word is a preposition, or a verb governing the instrumental case (which may also follow the noun).

c) Dative Case

This case may be ignored, since the generation of the English 'to' can be most economically handled in the dictionary listing of the manageable number of words which precede nouns in this case.

d) Nominative, Accusative, and Prepositional Cases

These may be ignored because of the factor of word order.

e) Number in Nouns

The plural number of all nouns is unambiguous, with the exception of neuter and feminine nouns in the nominative and accusative plural (where they are identical with the genitive singular). If these ambiguous forms have been identified as genitive (under la above), they may be automatically identified as singular also. In all other instances, the number of such forms can be satisfactorily determined by reference to the preceding word. The adjective and (in almost all instances) the preposition are absolute determiners of number; other forms which require the noun in the genitive case may also be utilized to determine the singular number of the ambiguous form (in instances where the English 'of' is not generated); the absence of these conditions, or the presence of a period or a comma in the preceding position, may be taken as an indication that the form is plural in number.

2) Adjectives

Often adjectives are useful in determining the case and number of nouns; otherwise, they may be ignored as to agreement with noun.

a) Short adjectives, singular, (in -zero, -а, -о) are to be preceded by the word "(is)" in translation; short adjectives, plural, (in -ы or-и) are to be preceded by "(are)". These English words are, further, to precede an adverb which may precede the short adjective.

Если температура очень высока
If the temperature (is) very high

b) Comparative adjectives: the word 'than' will be inserted in the translation if the following word is a noun.

3) Adverbs

The distinction between a short neuter adjective and an adverb is apparently impossible to make, since the forms are identical. Preliminary investigation shows that a high degree of accuracy can be attained by reference to context: if the following word is a modifier or a verb in the indicative, the word in question is an adverb; if the following word is an infinitive, the word in question is a short adjective. The accuracy of prediction can be increased by further extension of the comparison process. It is, however, doubtful that such refinement is necessary.

4) Participles

A participle may serve in a sentence as an "adjective", as a true participle or (rarely) as a noun. The decision as to its function in a given sentence cannot be made on the basis of form. Observation of its behavior, however, leads to the following formulation:

a) An active participle can be adequately translated as '-ing' Определяющий = 'determining'; a passive participle can be translated as '-ed' (определенный = 'determined').

b) If the participle agrees in case and number with the following word (a noun, or adjective + noun), it is treated as an adjective (i.e., as a modifier), число заряженных частиц = 'the number of charged particles' (rather than 'the number charged of particles').

c) If the participle does not agree with the following word, it is a true participle, число, определенное этим методом = 'the number, determined by this method.'

Again, although this formulation is completely arbitrary, no exceptions to its correct-

ness have been observed in a study of 132 occurrences. (Slightly less accurate results can be obtained merely by reference to punctuation: a preceding comma makes the word in question a true participle.)

The above represents the classes of syntactical problems which are encountered most frequently in Russian text. By application of well-defined rules involving reference to pre- or post-words, clarification can be attained to a very high degree of accuracy. A few minor problems remain, caused chiefly by "awkward" word order, inverted clauses, etc.

Conclusion: Syntactical ambiguity can be removed to a highly satisfactory degree by the comparison of ambiguous words with words in immediate contiguity.

Clarification of Semantic Ambiguity

It is obvious that problems of syntax and semantics are closely related. For purposes of discussion the two have been separated, and the latter has been arbitrarily divided into two categories: "structural" and "non-structural" clarification.

1. The most common instance of structural clarification is the determination of English equivalents by means of the grammatical case of contiguous words. Thus, the Russian preposition £ is translated as 'with' when the following noun is in the instrumental case, and as 'from' when the noun is in the genitive case. The English equivalent of other prepositions also varies with the grammatical case of the object, as set forth in dictionaries and grammars. These relationships are predictable and easily recognizable.

Behavioral analysis brings to light a great number of unsuspected semantic relationships between words of multiple meaning. These relationships have been only partially uncovered, but the semantic clarification so provided holds great promise in MT. An example is found in the Russian conjunction, и_, which is listed in dictionaries as: 'and', 'but', 'even', and 'also'. A test case was made of this frequent and annoying conjunction, on the assumption that perhaps its meaning could be determined by immediately contiguous words. On the basis of 200 occurrences in scientific text, it was found to be equated with the English 'and' whenever the preceding word was a noun (which situation prevailed in 70% of the total occurrences). By a slight extension of this comparison to other parts of speech and to punctuation, we can predict the correct equivalent of и in 90% of its occurrences.

Other examples of structural clarification of this kind include:

a) The word их. which serves in Russian both as a pronoun and pronoun-adjective ('them' and 'their' in English). It has been found that this word can be equated with the proper English word according to the nature of the following word (noun or non-noun).

b) Words which serve both as an adjective and as a noun, and whose English equivalent varies accordingly. Thus, данные is equated with ' given" when it is singular in number or when it agrees as a modifier with the following noun; in all other instances it is translated as 'data'.

2. "Non-Structural Clarification". Words of multiple meaning for which clarification by structural means is impossible constitute approximately one-third of the running words in a text. (This figure is in addition to idioms, which are a special problem.) In pursuit of the ideal — to select, within practical limits, a single correct equivalent for these words — we must look for some kind of contextual aid other than that supplied by grammatical features of surrounding words.

In the first place, it is clear that new techniques of lexicography for MT need to be developed. Reliance upon dictionary equivalents must be replaced by observation of the behavior of ambiguous words in given fields of technical writing. For example, if observation shows that the Russian изменение may be always equated with the English 'change', in texts on physics or mathematics, the nine equally possible dictionary variants ('alteration', 'fluctuation', 'variation', etc.) may be disregarded. Limited observation indicates that 'property' may be taken as the correct equivalent of свойство in the same field (as opposed to 12 dictionary listings); 'study' for исследование (7 listings); 'substance' for вещество (7 listings); 'body' for тело (8 listings); 'magnitude' for величина (15 listings), etc. In addition, superior techniques must be perfected for choosing the best "cover-word" from among a group of relatively synonymous equivalents. Existing "technical" dictionaries are in no sense idioglossaries, since they list a great variety of potential equivalents for most

words. A true idioglossary must be based upon the observed values of multiple-meaning words, with the emphasis placed upon singularity, rather than upon plurality, of meanings.

Regardless of the size of the context-sample, we must be able to observe ambiguous words in action: the kinds of nouns which follow certain prepositions, the kinds of adjectives which impart specific values to certain nouns, etc. An empirical study of this scope, practicable only with the aid of modern machine techniques, will go far towards unveiling the mysteries of "context". We have long since passed the stage in MT research when we should be bound by speculation of what "might be"; we need to take a bold step forward to find what actually exists.

The application of contextual analysis offers great potentialities for semantic clarification. In this instance, comparison of ambiguous words is effected with contiguous word classes. Word classes are simply groups of words (usually of like parts of speech) which have the common property of causing other words to behave in a predictable manner. For example, the Russian preposition по has ten potential equivalents when followed by a noun in the dative case; by reference to pre-determined noun classes we can reduce the number of choices to one, in most instances. (If the noun-object is an animate noun, по acquires the meaning, 'according to'; if the object is a verbally derived noun, the meaning is 'in'; if the object implies a path or a surface, the meaning is 'along'.) An extended survey of physics texts indicates that the vast majority of noun-objects after this preposition fall in one of these three classes. The word classes are formed purely on the basis of observed behavior; with further refinement and extension of research, it appears feasible that pinpointing of meaning will be possible for most occurrences of this most difficult preposition. Like procedures can be instituted for a great variety of ambiguous words.

The great advantage of using word classes is that the necessity of treating each new combination as an "idiom" is eliminated. It is apparently in some such fashion that the human translator chooses a particular equivalent for a given ambiguous word when he encounters the word in a novel or unremembered combination. In idioms, of course, the factor of memory proceeding from previous acquaintance with the combination, is essential. But when the human encounters the combination по оси for the first time, on what basis does he equate по with 'along' (the axis), rather than with 'in', 'according to', etc. ? It is possible that in some instances the human engages in a process of elimination, discarding from consideration certain inappropriate equivalents; it is also possible that the choice is often made purely on the basis of the "class" of noun-object (i.e., "axis" is associated with a class of words, including "line", "radius", etc., which is known, on the basis of previous experience, to impart the meaning 'along' to the preceding preposition). Just how decisive this type of word class association may be in the determination of meaning, and the extent to which the crudely formed classes described in the foregoing paragraph will answer the purpose, remains to be proved. It can safely be predicted that this kind of "contextual analysis" will be quite effective, particularly within specified areas of discourse.

Another type of ambiguity is posed by words which bear multiple meanings even within a specific area of discourse. The Russian noun напряжение, e.g., may be translated as 'tension', 'stress', or 'voltage'; it is obvious that any of these meanings may be applicable in a text on physics. A partial solution to the problem of choosing the correct equivalent may be sought in further refinement of the idioglossary: thus, in texts concerning electricity, 'voltage' may be predicted. The human translator often chooses 'voltage' because of the contextual aid provided by the subject area: specifically, he identifies the subject area by the title or beginning sentences of the text. Two mechanical methods may be adapted for determining the appropriate equivalents. One involves the employment of sub-idioglossaries (e.g., for the field of acoustics), — which may necessitate pre-editing, in texts which are not clearly or mechanically identifiable by subject area. Another possibility is the reference of multiple-valued words to certain key-words in the title or first sentences of the text. Preliminary study indicates that this approach may lead to unexpectedly positive results. To take an extreme example, it may turn out that the very presence of the word "polymorphic" in a title will fix the specific equivalent of the following polysemantic words in the succeeding text:

чистый	'pure', rather than 'clean', 'clear', 'net', 'smooth', 'absolute', etc.
твердый	'solid', rather than 'hard', 'tough', 'durable', 'stable', etc.
вещество	'substance', rather than 'matter', 'material', 'agent', 'composition', etc.
соединение	'compound', rather than 'fusion', 'connection', 'union', 'contact', etc.

(It should be noted that the fact that these words appear in an article on chemistry does not guarantee the same selection.) There may be no apparent reason that this selection of equivalents should be valid, and it is certainly possible to invent contexts within chemical literature where they would not be so. But, if on the basis of observation these equivalents are found to be adequate, there is a strong argument that the empirical evidence should be accepted and utilized.

There are, of course, words for which semantic clarification cannot be obtained by use of an idioglossary; the referent is not the subject area, but perhaps a contiguous word — an adjective for a noun, or a noun object for a verb. It remains to be seen whether or not the contextual aid provided by such contiguous words can be programmed in a non-idiomatic fashion, — i.e., not on a one-to-one basis. The goal should be the establishment of word classes of the "determining" words which will enable us to fix the semantic values of the "determinees".

The result of the aggregate of structural comparisons of this kind, and of the kind described in the preceding section, is, in effect, a new grammar — a structural, or analytic, grammar designed for the specific purposes of MT. There is no question that this approach, based on an analysis of ambiguous words in terms of coded features of contiguous words, is adequate for MT and is superior to the approach of conventional grammatical analysis.

From the point of view of methodology it is notable that a completely unexpected relation is found to exist between structural context and meaning. It should be stressed that the existence of this particular relationship has never been even remotely considered by Russian philologists. The connection is, of course, not absolute; it is merely one of the phenomena of language which can be discovered by observation, and which is sufficiently reliable to be of use in MT.

Conclusion: The value of contextual analysis for purposes of syntactic and semantic clarification should be evident. The plain fact, however, is that no systematic and thorough study of context has ever been attempted for any language. There is an overwhelming and immediate need for such a study, conducted over the range of a million or more running words in the scientific literature of a given language, with the help of machine techniques. The information and experience gained in such a study will be of great value for similar studies in other languages. Since our primary concern here is the behavior of words in context, the machine run should be constructed so as to give the researcher rapid access to numerous occurrences of ambiguous words in "real-life" situations. In line with Kaplan's suggestion, it may prove that five-word blocks (with the ambiguous word in the middle position) will be sufficiently large to establish semantic clarity and an adequate judgment of the effect of contiguous words.

A Refinement in Coding the Russian Cyrillic Alphabet

B. Zacharov, London University, London, England

By reducing the number of characters to be coded the problem of devising a numerical code for the Cyrillic alphabet can be simplified. This reduction can be achieved by providing code-words for only the lower-case forms of characters that do not occur initially; by disregarding the diacritic of the character ë, and by disregarding the character ë entirely. Ambiguities that arise in the latter cases can be resolved by an examination of the context.

THE PROBLEM of coding the Russian Cyrillic alphabet in numerical form has been considered previously in several papers [1] and it is clear that it would be desirable if each character of the Russian alphabet (together with any required numbers, punctuation marks and capitals) could be coded in such a way that a separate unique numerical code-word existed for each lower-case character, capital, etc. Unfortunately, the speed of modern digital computers and the size of their memories are such that a code of this form would result in considerable time being spent in the memory search for the appropriate target language equivalent.

It is clear, then, that ways must be found, apart from engineering advances, to speed up the memory search time. One way of doing this would be to decrease the amount of linguistic data stored in the memory, and this has been considered. [2] Another method would be to decrease the amount of numerical data (i.e., the number of bits) in the memory for a given number of source language characters. This last approach has been considered in a recent paper on mechanical translation[3] where all the lower-case characters, except ë, и, ъ and ь are represented by a five binary-digit code, while all the capitals and decimal numbers use a ten bit code; in the code proposed in that paper simplification is obtained on the basis of the statement that "... five of the 33 Russian letters never start a word and will not need to be capitalized ... ". The five Russian letters referred to are ë, и, ъ, ь, ы.

All the other Russian characters occur frequently in both upper and lower case and require to be coded separately in both these forms or by the same numerical code, except that the upper case is always preceded by some number which denotes an 'upper-case shift'.

Inspection of the statement quoted above reveals that it is formally incorrect with respect to ë although it is quite correct to state that none of the four characters й, ъ, ь, and ы ever begin a word in the Russian language so that clearly, it will never be necessary for them to be coded in upper-case form. (A rigorously phonetic transliteration of some other alphabet into Russian may create a trivial exception in the cases of й and ы This will not be considered here.)

---

1. Harper, K.E., "The Mechanical Translation of Russian: Preliminary Report", Modern Language Forum, vol.38, no. 3-4, pp. 12-29, Sept. - Dec. 1953.

2. Oettinger, A. G., "The Design of an Automatic Russian-English Dictionary", Machine Translation of Languages, John Wiley and Sons, New York (1955), pp. 47-65.

3. Wall, R. E., "Some of the Engineering Aspects of the Machine Translation of Languages", AIEE Transactions, I, vol.75, 580 (1956).

The Problem of ë

Reference to a Russian-English dictionary[4] shows us that many words of the Russian language begin with ë Notable examples are ёлка 'fir tree' and ёмкость 'capacity'; the latter is of especial importance in scientific texts.

Superficially, therefore, it would appear that ë should be treated in the same way as the other word-initial characters and that it should be coded in upper and lower case. However, the following points must be considered,

i) In practice, ë is never written in script form with the diacritic, either in lower or upper case — e and E are used.

ii) A modern standard Russian typewriter keyboard does not contain Ë or ë — the upper and lower case forms of e are used, as in (i).

iii) Both ë and Ë frequently appear in print, especially in the texts of scientific periodicals .

Thus, from (i), (ii) and (iii) above, it can be seen that the problem of encoding ë and Ë is complicated by the source of the Russian language text. If e and ë are coded separately, it would appear that words containing ë would have to be stored in the memory in two separate locations, with both e and ë in the corresponding positions of each word.

a) ë at the beginning of a word

For words with ë at the beginning, any coding difficulty can be overcome if it is noted that, if the diacritic is ignored, no ambiguity can arise. This is because no two words in the Russian language exist with different meaning such that corresponding letters of both words are the same except that ë at the beginning of the first word is replaced by e in the second word. As a result of this consideration it will clearly never be necessary to encode ë in capitalized form — the upper-case form of e will be sufficient.

b) ë in any letter position

If ë occurs in some letter position other than at the beginning of some word (x), ambiguity can arise only if another word (y) exists such that all the letters of the (y)-word are the same as the corresponding letters of the (x)-word except that ë in (x) is replaced by e in (y).

Examination of a Russian-English dictionary reveals that this does not occur often in the stem of a word. Similarly, experience tells us that ambiguity seldom arises as a result of word endings together with stem.

Examples of words where ambiguity may occur are:

все	all (plural)
всё	all (singular, neuter)
села	of the village (genitive, singular) / she sat
сёла	villages (nominative/accusative, pl.)

Whereas discrepancy need not necessarily occur in the first example, considerable ambiguity can arise in the second case since the words are different grammatical forms of widely different words (сёла is a plural noun while села may be a verb form or a singular noun).

However, we note that if the contexts of these words are examined, most cases of ambiguity disappear (this is especially true for Russian where strict grammatical rules concerning case endings and conjugation must be observed). Indeed, such an examination is essential for certain words in Russian and, more especially, in English. [5]

Certain Russian words are such that their spelling is associated with multiple meaning and, here, it is often the case that an examination of the context will not reveal which alternative is meant. In this event it becomes necessary to print out all the alternatives stored in the computer memory which correspond to the source word. At this stage a simplification may be effected if the computer dictionary is concerned only with a certain field (e.g., nuclear physics), in which case only those terms which may reasonably be expected to relate to that field will be printed out.

Examples of Russian words in such a category are:

замок	castle / lock
замотать	twist / shake

4. Smirinskii, A.I., Russian-English Dictionary, State Publishing House for Foreign and National Dictionaries, Moscow, (1952).

5. Yngve, V.H., "Syntax and the Problem of Multiple Meaning", Machine Translation of Languages, John Wiley and Sons, New York (1955), pp.208-226.

In the two examples above, ambiguity will disappear if the words are used in idiomatic context (e.g. padlock = висячий замок).

In the case of words containing e or ë, however, difficulties of multiple meaning that cannot be resolved by simple context (i. e., syntax) examination are very rare. In fact, in the author's experience, no example can readily be quoted.

Suggested Encoding Rules

From the above considerations, a set of rules can be formulated to include words containing ë and Ë. They are:

i) Source language words containing ë or Ë are stored in the dictionary in numerical form as if they contained e or E in the corresponding letter positions,

ii) Incoming source language words are coded with a unique number code for every lower-case character except ë which is treated as if it were e. All upper-case characters will have unique number codes corresponding to them (or they will be preceded by a coded upper-case symbol), except Ë, where the diacritic is ignored and the character is treated as if it were E; й, ъ, ь, and ы will have no upper-case code,

iii) If more than one target language alternative is found, the context of the Russian language word must be examined; this will also be required for any other word (not containing e or ë) where ambiguity may exist — as in the examples above.

The Problem of ъ

It may be noted that ъ could also be ignored completely since it occurs so very rarely in the Russian language. This may be of some importance since the character can be represented in several different ways, namely:

i) as ъ.
ii) as '
iii) as a gap in a word
iv) it is ignored completely.

As in the above encoding rules, if ambiguity occurs because ъ is ignored, the context of the word must be examined. An example of words where this kind of difficulty can arise is

сесть = sit down
съесть = eat

In these cases, if a unique meaning cannot be found simply from the program, all the target-language equivalents will have to be printed out and the required meaning determined by post-editing.

From an examination of the occurrence of e in the Russian language it seems that, if the diacritic is ignored the chances of ambiguity occurring in MT, with the rules formulated above, are very slight. Indeed, for a specific subject, where all the source language words in the dictionary are known, most cases of ambiguity and difficulties of multiple meaning could be overcome by sufficiently sophisticated programming techniques (i.e., syntactical and idiomatic context examination for all the cases of expected ambiguity).

As to ъ, it may be ignored in the encoding. The few cases of ambiguity will be resolved from a study of context.

Bibliography

A. F. R. Brown 122
Language Translation
Journal of the Association for Computing Machinery, vol. 5, no. 1, pp. 1-8 (January 1958)

A description of the method employed by the author in attempting to develop a program for the mechanical translation of 220 sentences from a French chemical journal into English is presented in this paper. The method is based on the formulation of a number of ad hoc rules that will serve to translate individual sentences as they are processed one by one. The rules are modified as new sentences are processed so that eventually a set of rules is developed that provides adequate translations of 90 or 95 percent of the sentences in a given text. The author also discusses and evaluates other approaches to the problems of mechanical translation. His conclusion is that the formulation of a number of small rules has more advantages than the formulation of a few very powerful rules because small rules can be more readily changed without necessitating the development of a new system. The questions of idiom translation; dictionary storage, and dictionary look-up are also considered.

J. R. Applegate

123
K. E. Harper, D. G. Hays, and A. Koutsoudas
A Glossary of Russian Physics on Punched Cards
The RAND Corporation, Santa Monica, Calif., P-1241, December 26, 1957

This is a description of a glossary of 6, 000 Russian forms, prepared by the University of Michigan and RAND, specifically for use in machine translation research and operation. These 6, 000 inflected forms of 2, 300 words have been punched onto IBM cards, each of which contains the Russian form, a form identification number, a word identification number, a syntactic code and one or more English equivalents. Duplicate sets of cards are available at cost to research workers.

E. S. Klima

124
H. P. Edmundson, K.E. Harper, and D. G. Hays
Studies in Machine Translation—1: Survey and Critique
The RAND Corporation, Santa Monica, Calif., ASTIA Document Number AD 150672 (Feb. 1958)

The paper presents a survey of results in MT. The problems are those of linguistics and those of automatic computation. In MT, linguistics deals with 1) meaning (word equivalence between the input and output languages, multiple meanings in either language, selection of words for the glossary, economy in storing) and 2) sentence structure (classification of words of the source language by functional type, etc.). Automatic computation, in MT, deals with the following problems: input (the perfection of automatic print reading devices for the varying states of the letters of a given alphabet and for a variety of alphabets), storage (requirements of large capacity on the order of 100,000, 000 bits, rapid access to any one of a large number of entries), computation (looking up words in the glossary, analyzing structure, resolving multiple meanings) and output (high legibility for direct process duplication for wide circulation) .

E. S. Klima

125
L. S. Barxudarov and G. V. Kolshanskij
K Voprosu O Vozmozhnostjax Mashinnogo Perevoda (On the Problem of the Potentialities of Machine Translation)
Voprosy Jazykoznanija, vol. 7, no. 1,(1958) pp. 129-133

Two linguists review the technological basis for automatic translation. They conclude that at present the realization of automatic translation is chiefly a linguistic problem. Although they stress the importance of formal structural analysis of languages, they caution that this approach cannot claim the role of a universal method for linguistic research.

A. T. Oettinger

I.I. Revzin 126
Addendum to the Paper "Some Problems in the Formalization of Syntax"
Bulletin of the Seminar on Problems of Machine Translation, Moscow State Pedagogical Institute of Foreign Languages, no. 3, (1957) pp. 20-29

Revzin reviews some work he presented at an earlier seminar, in the light of the approach outlined by Kulagina. The emphasis is on the application of set-theory concepts to problems of phonology.

A. T. Oettinger

127
H. P. Edmundson, D. G. Hays, E. K. Renner, and R. I. Sutton
Studies in Machine Translation — 5: Manual for Keypunching Russian Scientific Text
The RAND Corporation, Santa Monica, Calif., ASTIA Document Number AD 150668 (Dec. 1957)

It is a description of the methods used by The RAND Corporation for research in machine translation. The topics treated include: card layout, keyboard layout, punctuation (including special treatment of problems like capitalization), and the keypunch notation of special codes.

E. S. Klima

O. S. Kulagina 128
Ob Odnom Sposobe Opredelenija Lingvisticheskix Ponjatij (On a Method for Defining Linguistic Concepts)
Bulletin of the Seminar on Problems of Machine Translation, Moscow State Pedagogical Institute of Foreign Languages, no. 3, (1957) pp. 1-18

Four numbers of this mimeographed publication are known to have appeared, but only no. 3 was available for review in this issue of "MT."

The author points out that existing grammars are not adequate for automatic translation. She proposes the use of set-theory methods to develop an abstract framework based on precisely stated postulates. The mathematical development is relatively thorough, leading to the definition of sets corresponding, roughly, to paradigms, parts of speech, etc. The exposition is lucid, and has considerable pedagogical and heuristic value. In this paper, as in many similar efforts, the relation between the mathematical model and reality remains tenuous.

A. T. Oettinger

Paul Pimsleur 129
Semantic Frequency Counts
Mechanical Translation, vol. 4, no. 1/2, pp. 11-13

The success of a mechanical translation should be measured in terms of the level of depth required by the situation. To determine whether a careful translation is desirable a rough scanning will suffice. The use of cover-words, high frequency words that may be substituted for low frequency words, in the output language is an essential part of this process. The preparation of trans-semantic frequency counts resulting in dictionaries of reduced size that require less computer storage capacity is recommended.

Author

M. M. Masterman 130
The Thesaurus in Syntax and Semantics
Mechanical Translation, vol. 4, no. 1/2 pp. 35-43

The recent work of the Unit has been primarily concerned with the employment of thesauri in machine translation. Limited success has been achieved, in punched-card tests, in improving the idiomatic quality and so the intelligibility of an initially unsatisfactory translation, by word-for-word procedures, from Italian into English, by using a program which permitted selection of final equivalents from "heads" in Roget's Thesaurus, i.e. lists of synonyms, near-synonyms and associated words and phrases, instead of from previously determined lists of alternative translations. The Unit is investigating whether the syntactic properties of a word in a source language may be defined by a simple choice program, with reference to extra-linguistic criteria, which might be of universal or extensive interlingual application. It is hoped to combine or reconcile such a program with R. H. Richen's procedure for translating syntax by means of an interlingua. which has proved effective in a small-scale test. Studies have been made of the complementary distribution in literary English of words and phrases from "heads" in Roget, and of the construction of discourse from the contents of selected "heads". The possibility of producing a thesaurus better suited for machine translation purposes than Roget's, to be based on a more restricted lexis and a simpler categorization, is to be examined.

Author

Roderick Gould 131
Multiple Correspondence
Mechanical Translation, vol. 4, no. 1/2, pp. 14-27

It has been shown by Oettinger that the usefulness of rough Russian-English translations produced by an automatic dictionary is limited primarily by the large number of English equivalents which must be provided for many Russiar words. The design of an additional machine stage for reducing the number of equivalents requires that the words be somehow classified; this classification might be according to meaning, grammatical role in the sentence, or both. Detailed examination of a model automatic - dictionary output revealed that the multiple-correspondence problem arose primarily from nouns, prepositions, and verbs, in that order. However, the extremely small number of distinct prepositions involved suggests that they should be given special individual treatment. It is proposed that the "meaning words* (nouns, verbs, etc.) of Russian and English be classified according to meaning and the "function words" (prepositions, conjunctions, etc.) be omitted from consideration. Lists of meaning-class sequences appearing in large samplings of Russian text would be tabulated and stored in the translator; comparison with these tabulated sequences would then allow the number of different classes of English words corresponding to any given Russian word to be reduced.

Author

John P. Cleave 132
A Model for Mechanical Translation
Mechanical Translation, vol.4, no. 1/2, pp. 2-4

A mathematical model for a translating machine is proposed in which the translation of each word is conditioned by the preceding text. The machine contains a number of dictionaries where each dictionary represents one of the states of a multistate machine.

Author

R. B. Lees 133
Structural Grammars
Mechanical Translation, vol.4, no. 1/2, pp. 5-10

We adopt the view that the grammar of a language is a predictive theory which isolates the grammatical sentences of that language by means of immediate constituent analyses, morphophonemic conversions, and grammatical transformations. A sample grammatical analysis is given for the development of the verb phrase in German independent clauses. Simple rules are given for converting the verb phrase as a sequence of personal affixes, various auxiliaries, and the main verb into passive, future, or conditional clauses, and then introducing word boundaries, choosing the proper auxiliaries, arranging the word order, and finally mapping the resulting morpheme sequence into the correct sequence of words in the independent clause.

Author

Bjarne Ulvestad 134
Syntactical Variants
Mechanical Translation, vol. 4, no. 1/2, pp. 28-34

Traditional grammar is normally eclectic and vaguely formulated, and it often tends to overgeneralize or fails to state the range of validity for its rules. Grammars for mechanical translation must be all-inclusive and rigorously explicit. While the input language grammar must register all the grammatical constructions possible, the existence of basically synonymous morphological and syntactical variants permits considerable inventorial reduction in the output grammar. These considerations are discussed with reference to English and German examples: verb phrases with 'remember'/(sich)erinnern as the head; 'as if'/als ob clauses.

Author