[*Mechanical Translation*, vol.1, no.3, December 1954; pp. 41-46]

COMPUTING MACHINES FOR LANGUAGE TRANSLATION

T. M. Stout*
Schlumberger Instrument Company
Old Quarry Road, Ridgefield, Conn.

RESEARCH on the problems of machine translation has been going on for several years in this country and abroad. [1] To date it has been concerned primarily with the complicated linguistic problems involved in mechanical translation, since the engineers can probably build the necessary equipment. This article is intended to suggest some of the linguistic problems to the engineer and to explain some of the engineering ideas for the amateur or professional linguist. The reader is cautioned that the procedures and equipment described are not necessarily the best or most recent, and that considerable development must be done before an actual mechanical translator is built and put into operation.

General Approach: The Language Problem

Present proposals for a mechanical translator involve, in rough terms, constructing a machine which carries out automatically the process that the human translator is imagined to use in converting a sentence from one language (the input language) into a new language (the output language). This process is assumed to consist of (1) transferring the material from the printed page to the brain (reading); (2) searching a dictionary to establish the meaning or meanings of each word in the original text; (3) selecting the correct meaning from the possible alternatives; (4) rearranging and refining the results to fit the requirements of the output language; and (5) recording of the results in written or other form for future use. The general procedure may be illustrated by an example.

Suppose the translator is faced with the German sentence:

Er fand die Aufgabe zu schwer,

which may be translated, "He found the task too difficult." A German-English dictionary gives the following meanings for the individual words:

Er - he
fand (from finden) - found;thought,considered
die - the (article); that, this, he, she, it (dem. pronoun); who, which, that (rel. pronoun)
Aufgabe - task, duty; lesson, exercise; asking (of riddle); posting (of letter); registration (of luggage); giving up, shutting down (of business)
zu - to, at, in, on (preposition); too (adverb)
schwer - heavy; oppressive; clumsy; difficult; grave (illness); indigestible (food); strong (cigar)

Er can be translated only by "he." Although finden generally means "to find" in the sense of "to discover," it also has the figurative meaning, "to think" or "to consider." English "find" also shares these meanings and no great harm will be done if finden is always translated as "find." The presence of a noun following die, indicated by the capital letter or by a dictionary entry opposite Aufgabe, makes its translation "the." The translation of Aufgabe may be taken as "task" in all cases, since this meaning is general enough to include all of the other, specialized meanings; the nature of the task should be clear from the context. Zu is translated as "too" because of the following adjective, which presents the toughest problem in the sentence. The choice in this case evidently depends on the feeling that a task can be difficult, but not heavy, clumsy, grave, indigestible, or strong.

As this meaning suggests, a word which has only one meaning (or can arbitrarily be assigned only one meaning) will present no problems. Any

---

* This work was done at the Department of Electrical Engineering of the University of Washington in Seattle, Washington, and was originally published in THE TREND in Engineering at the University of Washington, Vol. 6, No. 3, p. 11 ff, July 1954. The author's interest in mechanical translation and many of the ideas contained in this article are the result of conversations with Dr. Erwin Reifler of the Far Eastern Department of the Univer-- sity of Washington.

[1] MECHANICAL TRANSLATION, Vol. I .March 1954, published at the Massachusetts Institute of Technology. An extensive bibliography of publications in this field.

word with several meanings, however, will cause considerable trouble. The selection of a particular meaning is sometimes based on grammatical considerations, sometimes on the presence of other words or types of words, and sometimes on the nature of the subject matter. In addition to the ability to read and write and search a dictionary, the machine - like the human translator - must be able to discern gramatical distinctions and the occurrence of words which determine the meanings of associated words.

Coding

At the present stage of development, it is assumed that the translating machine will work only with printed material. In addition to some obvious engineering advantages, this approach has the linguistic advantage that the written language is more distinctive than the spoken language. In English, for instance, the homonyms, not-knot, pair-pear-pare, and numerous other groups of words are easily distinguished by their spelling. The number of words with the same spelling and different pronunciations,such as lead-lead and bow-bow, is much smaller.

Since most computers are designed to work with numbers, the incoming text must be con verted from the written alphabet into a numerical form acceptable to the machine. Several different coding schemes are available for this purpose. One obvious procedure is simply to number the letters, using either two-digit deci mal numbers or five-digit binary numbers. Coded in this manner, A-B-C-D. . .would become 01-02-03-04..., or 00001-00010-00011-00100…

Other codes are commonly used in standard equipment which might be incorporated in a translating machine. Machines available from IBM use the code given in Table I, in which each letter is represented by two holes punched in a column of a standard punched card; the upper hole is called a zone punch and the other is a digit punch. Standard teletypewriters use the Baudot code given in Table II, which employs five pulse positions in a manner similar to the binary code (plus a sixth pulse for timing).

Binary or teletype coding requires more digits for each letter than the decimal or IBM coding and might appear to require considerably more space. On the other hand, these codes employ only two symbols (0 and 1, pulse and no pulse) for each digit. The physical elements in the computer can therefore be simple two-state devices, such as a switch or relay whose contacts are either closed or not closed, a vacuum tube which does or does not carry current, a magnetic core which is magnetized or not, and so forth. Since it is easy to determine which state exists, reliable operation is obtained without any accurate measurements or precision components.

TABLE I
ALPHABET CODING USED IN IBM PUNCHED CARD EQUIPMENT

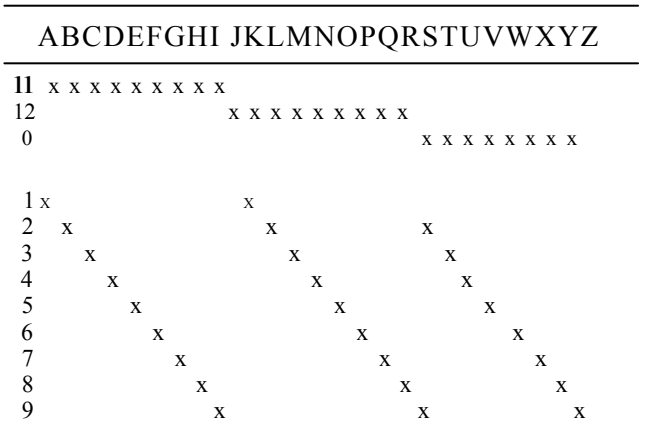

TABLE II

STANDARD BAUDOT TELETYPE CODE

LETTER	PULSE 1 2 3 4 5	LETTER	PULSE 1 2 3 4 5
A	X X	N	X X
B	X X X	O	X X
C	X X X	P	X X X
D	X X X	Q	X X X X
E	X	R	X X
F	X X X	S	X X
G	X X X	T	X
H	X X	U	X X X
I	X X	V	X X X X
J	X X X	W	X X X
K	X X X X	X	X X X X
L	X X	Y	X X X
M	X X X	Z	X X

Input and Output Devices

A number of standard devices are available for coding the incoming text for insertion into the machine and, after the translation process is completed, for decoding and printing the translation in the output language. Teletypewriters, operated by typists with no knowledge of either language, could be used to supply

electrical signals directly to the translating machine or to prepare punched paper tape for later use. Similar machines can be used to type the final output of the translator.

Input devices now available are relatively slow, so that faster means of supplying material to the translating machine would be essential. An electronic reading device, capable of working directly from the original printed text, has recently been announced.[2] Faster output devices will also be required to maintain over-all balance.

Storage

The dictionary needed in a mechanical translating machine might be stored on a magnetic drum such as the one shown in Fig. 1. This type of storage, in which information is stored by magnetizing small areas on the surface of a revolving cylinder, is widely used in arithmetic computers and has a number of desirable properties: a large ratio of information to volume, lower access time, permanence, and simplicity.

Individual words are stored along the length of the drum (each letter being represented by a group of five magnetized or unmagnetized spots) and pass the reading heads once in each revolution of the drum. Words in the input language are stored at one end of the drum, and their equivalents in the output language at the other end. If the drum is rotated at 2,400 rpm, or 40 rps, each word is available in not more than 25 milliseconds. Following standard practice, 80 spots per inch can be placed around the circumference of the drum and 8 tracks per inch along the length of the drum. Allowing 10 letters or 50 tracks per word in both halves of the dictionary, a drum 12.5 inches long and 12 inches in diameter would hold approximately 3,000 words and their translations.

In order to reduce the average time spent in searching the dictionary, certain common words might be stored several times on the same drum. The 850-word vocabulary of Basic English could be stored three times on a single drum, so that any particular word is available

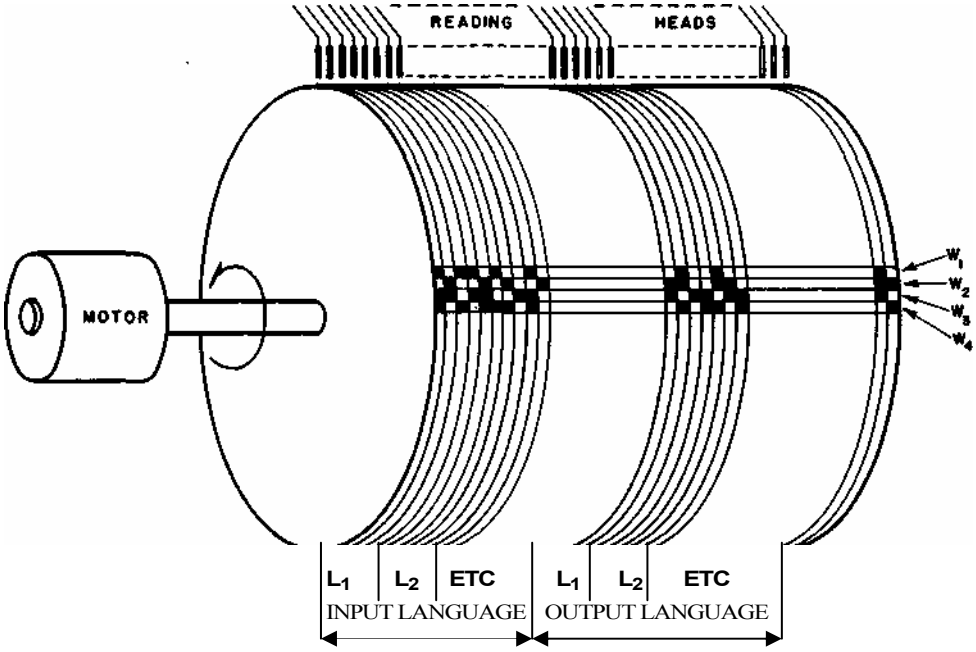

FIG. 1. MAGNETIC DRUM FOR DICTIONARY STORAGE

Words (W_1, W_2, etc.) are stored along the length of the drum,and each letter (L_1, L_2, etc.) requires five tracks around the drum.

2 Shepard, D. H., "The Analyzing Reader." A paper presented at the IRE convention in San Francisco, Aug. 19, 1953.

in a third of a revolution or less (not over 8 milliseconds).

To provide an adequate vocabulary for satisfactory translation, several such drums would be required. By searching all drums simultaneously, as explained below, any word in the dictionary could be found in the time required for one drum revolution. At approximately one cubic foot per drum, exclusive of the associated circuits, the space required for a vocabulary of 100,000 words or so becomes rather large. A number of tricks are available, however, for reducing the size of the mechanical dictionary.

If we are concerned with translation into English, as seems probable, many words in the input language text will not require translation. English has borrowed extensively from other languages and many foreign words are immediately recognizable by the English reader. A glance at a German dictionary, for example, reveals such words as Deck, Despot, Diplomat, and Dock which are identical with the English forms; we also find Demagog, Demokrat, direkt, Distanz. and Doktor which differ slightly in spelling but would present no real difficulties to the reader. The translation process can be by-passed for such words, and the original input word printed directly in the output. This approach must be used with caution, since the two languages may not share all the meanings and connotations of a given word, but it does offer hope for tremendously reducing the size of the mechanical dictionary.

Compound words are rather common in German and can, in fact, be invented at will by writers and speakers. If the meaning of a compound is clear from the meanings of its constituents (as is likely for all except old well-established compounds, which will be entered as distinct words), the dictionary can be searched for each constituent separately, and the respective translations compounded on the output side.

Endings, used extensively in other languages to convey grammatical information such as tense and number, can be treated in similar fashion to effect a further reduction in the size of the dictionary. Each word might be regarded as a compound built from a stem, common to all forms of the particular word, and an ending, which may be shared with other words. The dictionary may then be split into a large stem section and a small ending section. A useful by-product of this procedure is the grammatical information made available by the identification of an ending; this may be used in the elimination of impossible translations, discussed hereafter.

The techniques used in the dissection of compounds will be valuable in still another way. If a word has more letters than are permitted by the physical size of the dictionary (ten letters in the example above), it can be split into two parts which separately signify nothing. Beratschlagen. for example, might be split into Beratsc and hlagen, with parts of the translation stored opposite each half. Dictionary space is used more efficiently in this manner, but the processing time may be increased excessively.

Splitting words in order to determine parts of a compound, or stems and endings, is fraught with difficulties which must be explored by linguists. The engineering techniques for carrying out these operations have been devised, but are too involved to discuss here.

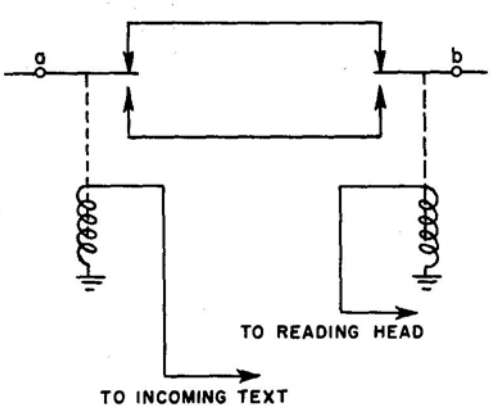

FIG. 2. RELAY CIRCUIT FOR CHECKING ONE PULSE OF INCOMING TEXT WITH ONE TRACK ON DICTIONARY

Dictionary Search

In making a mechanical translation, the first step is a comparison of each word of the incoming text with the entire dictionary. If any word is not found in the dictionary in its original form, the dissection scheme for endings and compounds can be tried; if this fails, the word can be printed through without alteration.

Several methods are available for making this comparison; an impractical but easily understood system is shown in Fig. 2. This system requires two single-pole double-throw relays for each pulse position: one relay operated by the incoming text and the other relay operated by pulses from the reading heads on the magne-

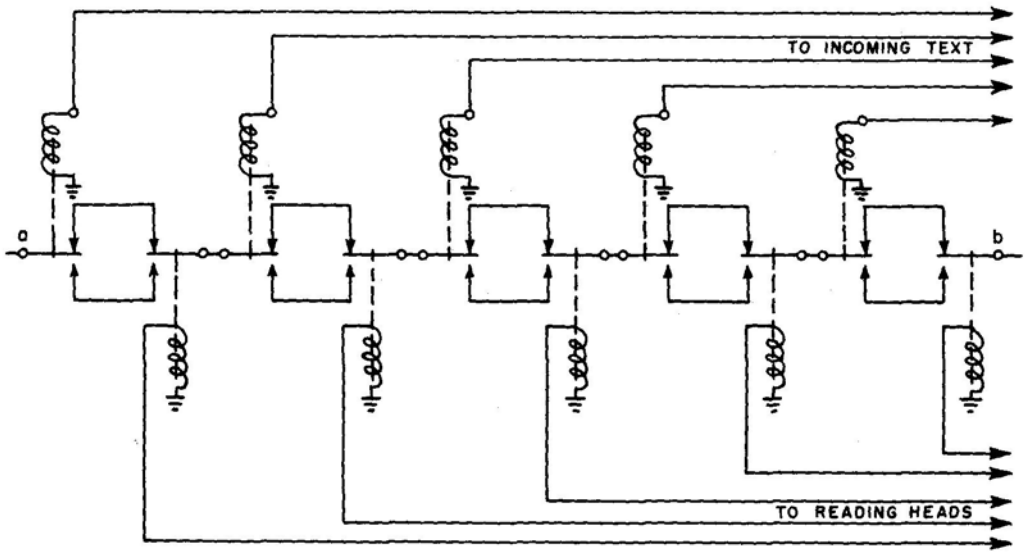

FIG. 3. COMBINATION OF RELAYS FOR CHECKING ONE LETTER

tic drum. The path between points "a" and "b" is closed only when both relays are either energized (pulses present in both incoming word and dictionary) or not energized (spaces present in both places). The occurrence of a closed path, therefore, indicates that the particular pulse position is identical in both the incoming word and the dictionary.

Entire letters, coded as a group of five pulses or spaces, can be checked by a series combination of five such relay circuits, as shown in Fig. 3. In corresponding fashion, words of ten letters could be checked by a series combination of fifty such relay circuits. A closed path through a long string of such circuits indicates that the incoming word has been found in the dictionary, and this event can be made to initiate printing of the translation stored at the other end of the drum.

An input-language word with several meanings can be entered in the dictionary several times, each time with a suitable translation. The searching procedure outlined above would uncover each of the possible translations and would make them all available for further consideration. To assist in the subsequent selection of one of these meanings, each translation might have a "tag" stored with it, which would supply grammatical or other necessary information needed by the machine.

With a multiplicity of such circuits, a number of dictionary drums could be searched simultaneously, as suggested schematically in Fig. 4. The incoming text is supplied to all drums at the same time. Correspondence between the incoming word and a dictionary entry is noted on only one drum, from which the translation is obtained. Parallel operation of this type would permit a dictionary of any desired size with the access time of a single drum, but at a considerable price in additional checking circuits.

In a practical comparison system crystal diodes, transistors, or vacuum tubes would be used instead of relays. These elements have no moving parts to limit the speed of operation and require much less signal power.

Multiple Meaning

Having obtained the possible translations for each word in a sentence, the machine is faced with the problem of selecting the correct meaning from several alternatives. This problem can be attacked in a number of ways.

In technical writing many words have specialized meanings which are used in all texts in a given area of science. For example, Flügel in a paper on aeronautical engineering is much more likely to mean "wing" than "grand piano," both of which are given in a general dictionary. The machine could be instructed to select the specialized meaning when the text is known to be in a specialized area (by means of appropriate tags) or special dictionaries could be used.

A number of distinct problems can be recognized in the case of general language. As indicated by the examples, the translation of a word

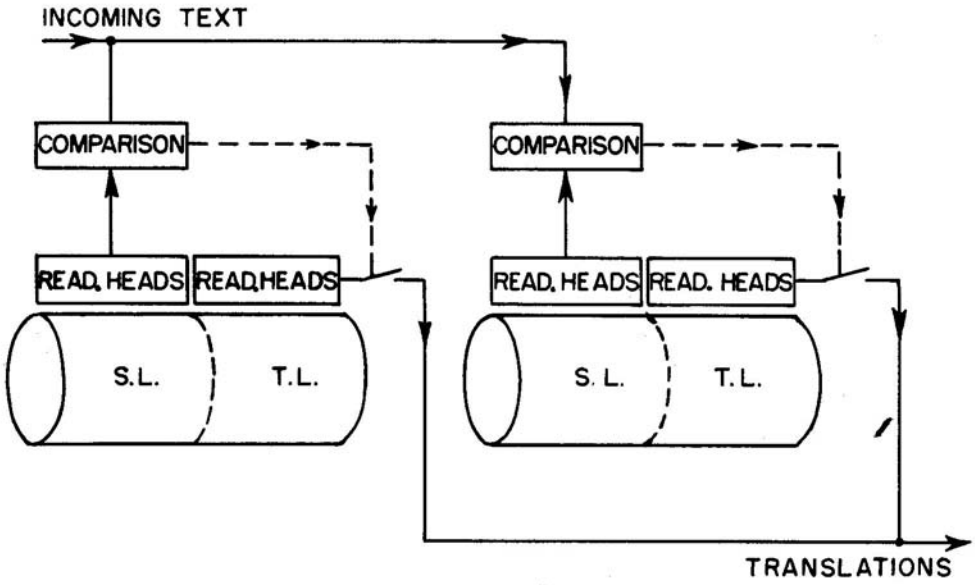

FIG. 4. SCHEMATIC DIAGRAM SHOWING SIMULTANEOUS SEARCH OF TWO DICTIONARY DRUMS

is sometimes based on grammatical considerations, sometimes on the co-occurrence of another word or type of word in the same sentence or clause, and sometimes on the larger context. In all cases, the choice is determined by examining the surrounding words and, according to rules furnished by the linguists, either selecting or eliminating certain alternatives.

The general procedure employed by the machine in selecting the proper meaning can be indicated by an example. For the German sentence given above, a superficial study suggests the following rule for the translation of zu: if zu is followed by an adjective or adverb, its meaning is "to," but otherwise it is a preposition, and its meaning must be determined by additional analysis. The translating machine can be instructed to examine the tag on the word following zu and, if the code designation for an adjective or adverb appears, to select "too" as the meaning.

Not all words present difficulties with multiple meanings, and the mechanical translator can easily locate the trouble-makers in any sentence by counting the alternatives encountered in the dictionary search. Having found a word with several possible meanings, the machine can refer to a list of rules appropriate to this word or its general class of words. This list should be flexible, so that rules can be added or discarded without disrupting the operation of the other rules. The machine can probably be arranged to count the number of times each rule is used and the number of successes scored, so that the effective rules can be applied first and ineffective rules discarded.

The linguistic rules will necessarily be coded and could, in fact, be expressed in algebraic fashion by the techniques of symbolic logic.[3] The resulting algebraic expressions can be simplified by formal procedures and can be converted directly into devices which carry out the selection process. The so-called logic circuits needed in a mechanical translator are employed in conventional arithmetic computers and their design should pose no special problems.

Conclusion

Experiments with word-by-word translation by mechanical means have already been conducted with surprisingly good results, even where no attempt has been made to deal with the problem of multiple meanings. With even a rudimentary set of rules for selecting or eliminating some of the possible meanings, still better results should be obtained. If the linguists can discover the rules, the engineers are ready to build the equipment, given the necessary support. Practical mechanical language translation is a definite possibility for the near future.

3 Langer, S. K., AN INTRODUCTION TO SYMBOLIC LOGIC: New York, Dover Publications, 1953.