THE USAF AUTOMATIC LANGUAGE TRANSLATOR, MARK I

George Shiner
Rome Air Development Center
Griffiss Air Force Base, N. Y.

Summary. After consideration over several years of many forms of equipment in which automatic language translation could be effectuated, it was decided that the practical and economical scheme was a mechanized dictionary capable for the storage of all the words in the foreign language and the English equivalents. Additional high speed electronic circuits and equipment could be electrically connected to syntactically and semantically process the dictionary output for a more accurate and intelligible translation. The essence of the USAF Language Translator program is that it is based on a large-capacity, rapid-access photoscopic memory to be used as the dictionary. The memory is capable of storing thirty million bits of digital information with a maximum access time of 50 milliseconds, and has a scanning rate of (l) million bits per second. The actual area for storing the thirty million bits is only ten square inches. Thus, the memory has a density of three million bits per square inch. Utilizing this memory, it is possible to look up or translate twenty words per second, or to search the entire memory or dictionary in thirty seconds. This paper discusses the Rome Air Development Center program for automatic language translation based on the capability of the aforementioned device. Some technical characteristics are provided and illustrated.

Why Automatic Language Translation?

1. There are Air Force and other Government Agencies which are concerned with the fact that scientific, technical, and cultural knowledge is recorded in a multitude of languages and fields, so that no one person has easy access to it all. This state of affairs is very costly in duplication of effort in science and technology, not to mention the lack of understanding in cultural matters, the lack of communication and understanding among nations.

2. Many efforts have been made to alleviate this situation. For one reason or another, none have been very successful. A large amount of the world's printed output remains untranslated, and a large portion of the worthwhile material of interest to the Air Force remains untranslated. Difficulties include the fact that a human translator's vocabulary is limited, the cost of human translation is high, and qualified technical translators are difficult to obtain. We are forced to automatic language translation as a way through this barrier of international communications.

3. It has been conceived that perhaps an electronic digital machine could be used to help with the translation of languages. This probably seems to be a fairly naive statement after reading recent publicity on automatic language translation. The fact is that the many well publicized attempts on automatic language translation fall far short of the total task and fail to give the real and complete story. Serious workers in the field of automatic language translation are first to admit the great difficulties in the way of an acceptable solution. However, there seems to be very good reasons for believing that an acceptable solution can be found. The output of first attempts at an automatic language translating machine may be no literary masterpiece, but it should be satisfactory for many purposes. For example, it might be adequate for the use of the scientist or engineer tapping the tremendous talent that expresses itself in foreign literature in his field.

4. Much of the problems of keeping up with the literature is concerned with looking over articles in a rather cursory manner and deciding which ones merit more careful attention. For every important article, there are usually many that are unimportant for that particular reason. Thus the engineer can scan and discard 100 documents by seeing only a rough translation made by a machine, and can select one in which he is particularly interested. Eventually even the necessity for refined translation by human translators may vanish, since it is highly likely that experiments with the first machine may give data to modify and improve this first machine leading to a second machine, until a machine is achieved that will give a translation that is accurate for almost any purpose.

5. There are many less obvious but important results of automatic language translation. Most important, however, is that mechanical translation of foreign scientific and cultural writings *is* bound to have a great effect upon Air Force Research and Development, with eventual great impact on international communication and understanding; on our own culture, science and technology.

Automatic Language Translation Requirements

Elements of machine translation

1. The dictionary corresponds to what is

called the memory in a digital computer; we must store all the words of the dictionary in the memory of the automatic language translator which must have all source language words. The source language is the language to be translated. In this particular development the source language is Russian. The target language is the language into which translation is made, and in this case it is English. Each source language entry must have the necessary corresponding target language equivalents stored with that entry in the memory.

Memory Requirements

1. Linguistics has shown that even words of rare occurrence such as new and ultra-technical words in space technology must be translated, since these words usually are the key to the new idea expressed in the sentence. Therefore, it is necessary to store a complete source-target dictionary, with the possible exception of internationals which are transliterated. Webster's New International Dictionary, second edition, contains some 550,000 vocabulary entries. The Russian language may not have as many words as English, but the storage requirements are about the same because of the Russian use of varied word forms to denote the grammatical usage of a word.

2. Previous publicized attempts at automatic language translation, although informative, did not have any practical applications because of the enormous memory requirements of the dictionary. There are linguistic and computer programming schemes that decrease storage requirements, but they all present logical difficulties in their complete form, and represent a fundamentally expensive approach to automatic language translation.

3. A pertinent factor in using optical memories is its great economy. It is estimated that the cost of translating one Russian word using an optical storage will be approximately five-hundredths of a cent, compared to one-fourth of a cent per word using magnetic tapes and other necessary associated equipment.

An Automatic Russian-to-English Language Translating System

Language Automata

1. The first automatic language translation equipment in the United States is being built for the USAF and is near completion. The Research & Development is being sponsored by Rome Air Development Center, Griffiss Air Force Base, New York. It will automatically translate Russian into English and is designated the USAF Automatic Language Translator, Mark I.

2. It will not be necessary to know or understand the Russian language to translate with the Mark *I*, Of course, a clearer understanding of the subject material translated by the Mark I will be obtained by having knowledge of the translated subject.

3. Not only is Mark I reliable since its memory cannot accidentally be changed or erased but it allows translation at any time and at any location. The only advanced preparation for obtaining automatic language translation is to place the source language on punched paper tape. Once the paper tape is prepared it can then be shipped to the various facilities for translation by the Mark I, or the information on the paper tape can be handled over proper communication channels for remote and instantaneous translation. (See Fig. 1)

Input Operation

1. The automatic language translation process starts with typists who are familiar with only the Cyrillic characters of the Russian alphabet. The typists use a Flexowriter to punch the text to be translated, to produce the paper Input Tape which is required for the input to the translating system. The Input Flexowriter is essentially a standard Flexowriter machine but with a few code changes and the keys labeled to correspond to both the Russian and English alphabets, Arabic numerals and punctuation signs. (See Fig.2)

2. With Cyrillic type faces on the Input Flexowriter it is also possible to type out a copy corresponding to the input material. This will be useful for proof reading or later comparison with the translated output.

Automatic lexicon

1. The lexicon in the Mark I is on a large-capacity, rapid-access photoscopic disc. It serves as an automatic dictionary in that correspondences between the source language and target language are stored there. The source entries are semantic units of Russian or word blocks. Each inflected form of a word is a separate entry. Hence, each form of a word is not found by splitting of the endings and looking for the stem. This aids materially in providing grammatical structure and in the solution of the multiple-meaning problem. Although the storing of each inflected form of a word as a separate entry increases the number of entries by more than an order of magnitude, this is not a problem in a large-capacity, rapid-access store; and it simplifies the search logic.

2. The structure of the entries and the search logic is desired not only to find the target equivalents of a word, but also of idioms or semantic units of more than a word.

3. When the Input Tape is fed into the Mark I several processes take place. Russian input is compared with the contents of the lexicon in the Photoscopic Information Storage Unit and translated by semantic units to produce an English output.

4. The search procedure is similar to the method that a person uses in looking up a word in a conventional dictionary. A track on the disc

corresponds to a page in a dictionary. Each track has about 250 entries, each beginning with a Russian semantic unit. The alphabetical symbols *of* the unit have been assigned binary codes which allow the words to be arranged in order. Certain punctuation marks have semantic meaning, so these symbols, as well as the space between words, are also treated as letters.

5. It should be emphasized, that the Mark *I as* currently designed will supply not only the meanings or multiple meanings *of* the semantic units, but also, with the aid *of* editing symbols prefixed to the meanings, the grammatical data may be more elaborate than the schoolboy's eight parts of speech. *If in* addition, a data-processing unit, which could be called a syntactical processor, were included in the system to which the large-capacity memory supplied the rules for operating on the syntactical data, the context of the whole sentence could be used to arrange the words in the proper order, and increase the probability of selecting the proper meaning from the possible multiple meanings.

6. In order to use a translating system based on syntactic operations, even for one source and one target language, it is necessary that a great deal of research and development be done to determine the syntactic data.

7. The research could be performed faster, simpler, and more efficiently with the aid of the Mark I Translator. The machine can be put to work and introduce an iterative or feedback process from the Mark I output to semantic and syntactic data.

Transliteration

1. If a word is not found in the lexicon, the Russian word is transliterated and typed in English characters on the Output Flexowriter. Transliteration is automatic and will cause the color-shift key on the Output Flexowriter to function so that transliterated words are printed in red ink. This will be useful in picking out words not yet in the lexicon, or new words that the Russians are using. The discovery and knowing the frequency of new or certain Russian words can be very useful.

2. Non-Russian input such as numerals is also sent directly to the Output Flexowriter. Thus the output from the Mark I consists entirely of the conventional English letters, numbers and punctuation signs.

Output

1. The output (English words) is produced with changes to facilitate the interpretation of the transliteration. The output translation will have the same punctuation symbols as the Russian input text. In addition to these symbols, certain editing signs such as the slash and asterisk will be used for multiple meaning interpretation. In addition to these editing signs, subscript Arabic numerals can appear after certain words with multiple meaning. Subscripts following a word may be used to identify the technical fields in which the word has *a* specialized meaning. Subscripts preceding a word may be used to convey some syntactical data which can be helpful in pinpointing meaning.

2. Provisions have been made in the Mark I to main sentence structure correctly on the Output Flexowriter. The paragraphs on the Output Flexowriter start at the same place in the text as the Input Flexowriter. Carriage return at the end of lines, within paragraphs on the output, will also be controlled automatically.

Description of USAF Automatic Language Translator Mark I

Photoscopic Information Storage Unit

1. The Photoscopic Information Storage Unit currently being built is near completion and may be briefly described as follows. It consists of a flat glass disc, 3/8 inch thick and 10 inches in diameter, on which information has been recorded photographically in 700 concentric tracks in an annulus of 0.36" width. The information is recorded in tracks of binary digital form, as black and white marks, 350 micro inches in width and height on High Resolution Photographic Emulsion.

2. In recording the information digitally on photographic emulsion, a double redundant code (white-black for 0, black-white for 1) is utilized for technical reasons (elimination of large black or white areas where chemical development would be uneven, simpler photo-multiplier tube detection circuits, establishment of a gray level for the servos, etc.) (See Fig. 3.)

3. The disc is continuously rotated at 1200 r.p.m. by a reading station. This reading station consists of a cathode ray tube as the light source a microscope objective lens projecting an image of a spot from the cathode-ray tube face onto the information track, and a photomultiplier tube immediately behind the disc, to detect the light transmitted by the black and white marks in the tracks as they pass by. Deflection of the electron beam in the cathode-ray tube allows the spot to jump to any one of the 700 tracks in the 0.36" annulus. The low inertia in switching this "reading head" allows a two-dimensional search over the thirty million bits in the 10 sq. inch annulus, so that essentially random access can be achieved in the time of one revolution, or 50 milliseconds. At a nominal speed of 1200 r.p.m. and marks 350 micro inches wide, information is read at the rate of one million bits a second. The access time can be reduced by an order of magnitude by increasing the speed of rotation of the glass disc or the number of reading heads. (See Fig. 4)

4. The output current from the photomultiplier tube follows the succession of black and white squares. This electrical signal is fed to

computer-type circuits, and compared with the words in an input register, which is continuously loaded from the Input Tape. In a match, the subsequent information is delivered to an output register. Here it is temporarily stored until delivered to suitable output equipment such as a high speed line printer.

Recording System

The Tape-to-Film Transfer Unit

1. In order to get information into the memory unit, i.e., photographed on the disc, from an external source, a recording system has been designed. (See Figure 5.) The Recording System consists of two basic units.

2. In the Tape-to-Film Transfer Unit, the first step is to get the information on to punched paper tape. The process starts with a collection of dictionary entries prepared by linguists specifically for the automatic language translation process. These entries consist of Russian semantic units (words or group of words) together with control symbols and the associated English output translations and editing symbols. The entries are then punched on IBM cards using a code developed to keep within the limitations of standard IBM equipment. With the entries on these cards, it is easy to tabulate, arrange, augment or process the lexical information as desired. When the lexicon is complete, the IBM cards are then sorted in the order required for entry on the photoscopic disc. The cards are then employed to print out an IBM tab sheet for proof reading. The next step is to prepare a Disc-Loading Tape from the IBM cards with standard IBM machines and a special machine - the Tape-to-Tape Unit - which converts the information into 6 hole punched paper tape.

3. The disc-loading tape containing the lexical entries is then fed into the Tape-to-Film Transfer Unit, designed specifically to arrange the material in proper coded form in tracks on 70MM photographic film. The purpose of the Tape-to-Film Transfer Unit is to transfer the information from the punched tape to the film in exactly the form to appear on the disc. By "form" is meant the correct binary pattern, control marks, and borders for tracking. However, the form is not geometrically identical. The information is most conveniently recorded on the film in linear tracks along the length of film. One such linear track is to correspond to a circular track on the disc. A roll of 70-mm film will have 77 such contiguous tracks, corresponding to the 77 continuous concentric tracks on the disc. Nine rolls of 70-mm film will fully load the disc. The arrangement of the information on the film is exactly as it will appear on the disc.

4. The general procedure is then to read a 6-bit character from the punched tape reader. These 6 bits are fed to a buffer store from which they are read serially, to gate a light source. The succession of flashes, 12 per character, are photographed on the 70-mm film which is moving smoothly in front of the light source. Everything on the punched paper tape gets transferred to the film. In order to have uniform mark sizes, the timing of the flashes, and the velocity of the film, must be precise.

5. A nominal speed of 200 characters per second (1200 bits per second) has been chosen as the input speed of the punched paper tape. Each track, hence each roll of tape, has nominally 7200 characters (43,200 bits of information), and corresponds to one circular track of the" same capacity on the photoscopic disc. The 77 tracks thus contain approximately 5.5 X 10 5 characters of 3.3 X 10^6 bits. Nine rolls of 70-mm film will be needed to load the disc fully, with 5 X 10° characters, or 30 X 10^6 bits of information.

6. One 43,200-bit track is photographed on the 70-mm film in 45 seconds. Another 45 seconds are required for rewind, and 30 seconds for changing tape. A roll of film can be filled with 77 tracks in less than 3 hours. A roll of punched paper tape, used in making one track on the 70-mm film, is 150 feet long, with 15 feet of leader and trailer.

, The Film-to-Disc Transfer Unit

1. The other basic unit for recording information on the disc is the Film-to-Disc Transfer Unit. The information is recorded on the disc by photographing a display, consisting of an illuminated negative, namely, a roll of 70-mm film made in the Tape-to-Film Transfer Unit. Successive photography of nine rolls is required to load the disc fully.

2. There are two basic operations to be carried out in the Film-to-Disc Unit: (l) to reduce the size of the pattern some sixty fold, (2) to transform the linear array into a circular one.

3. As stated before, transfer of information on the photoscopic disc takes place in two steps: the first is to 70-mm film and the second is from the 70-mm to the photoscopic disc. The arguments in favor of making the transfer in two steps are as follows: The code was chosen for optimum operation of the Disc Reader, and does not correspond to any conventional punched card or tape coding. Hence, the coding has to be changed. It does not seem desirable to make this rather complicated change of code in a single operation of transposing information from an outside source directly to the disc. Also, it is advantageous to have all the coded information in a final form or display before the final operation. In addition to the above advantages, this two-stage procedure allows, for easy preparation of display, a reasonable size in which the

marks are already small, so that excessive minification on to the disc is not required. A sixty-fold reduction is quite practical. Also, by photographing a large number (77) of tracks simultaneously, registration is avoided in the regime of small dimensions on the disc. The photograph of one roll of film on the disc produces a band of 77 tracks. Nine rolls of film produce nine bands which comprise the total annulus of recording on the disc. Registration of the successive bands on the disc need not be unduly high as the allowance between them is only a small percentage of the total store on the disc. This is to be contrasted with the much smaller tolerances which would be required if tracks were laid down one by one on the disc*

4. The conversion from rectangular to annular form is not difficult because the effective radius of the disc is large (300") when the image, or disc, plane is projected on the object, or film plane. A factor in the geometrical transformation from rectangular to annular form is the relative velocities of the film and the disc. When successive films are photographed by shifting a lens radially, there is a differential velocity of the film with the emulsion on the disc because the linear velocity of the emulsion on the disc varies with the radius. To compensate for the differential linear velocities, the film is tilted under the optical axis so that its velocity components in the optical plane compensate for the change in the emulsion velocity due to change in radius.

5. The Film-to-Disc Unit consists of a film transport component or projector which moves the 70-mm film through a gate. Here the film is illuminated and a reduced image is made by a microscope objective on to the disc. The disc is rotated so that one revolution is made in the same time that the roll of film passes through the optical gate. After this photography, a new roll of film is put in the projector and the lens moved inwards radially* The disc rotated again and the second band recorded on the disc. This process is repeated until the nine rolls of film forming nine bands in juxtaposition have been photographed on the disc. The disc is then developed, washed, and dried and is ready for use in the Disc Reader.

6. The photography of one 150 ft roll of 70-mm film, to make one band on the disc, takes one minute. The recording of nine bands of 30 *X 10°6* bits of information will take nine minutes, neglecting film reload time.

7. The digital information is recorded on a 10-inch diameter optically flat glass disc. The nine bands of information form a concentric annulus 0.36 of an inch wide contain 700 concentric tracks of information. Thus the digital information has been recorded on the glass disc at a density of 3 X 10 bits per sq. inch. Figure 6 is a photograph of a completed photocopic memory. Figure 7 is a drawing of the glass disc.

8. This recording system, the Tape-to-Film transfer unit and the Film-to-Disc Transfer Unit has been constructed and completed. Discs which are virtually free from defects can be made but complete consistency of quality is lacking. It is presently being tested and finalized for developmental use.

Future Possibilities of Photoscopic Storage

1. Progress with the Photoscopic Storage Information Unit now being built for USAF has progressed far enough that it is certain that it can be made to work. Information has been gathered to show the possibilities of further development of this type of memory device.

2. The interesting feature of the present system is that no design factor is being pushed to *a* limit, so that it should be possible to obtain an order of magnitude improvement in capacity, access and reading rate.

3. The size of the marks is now one-third of a thousandth of an inch. Marks one third of this have been photographed—which means a tenfold greater density is possible, as far as making the disc is concerned. The ultimate limit is the resolution of the lens, not the graininess of the emulsion. The reading of marks at a higher density will require more light. At least ten times more light is now available. The smaller marks will also require greater precision in keeping the focal plane of the reading lens on the surface of the disc. This may require an air bearing or a focusing servo system, but it is certainly feasible. Higher reading rates or shorter access times can also be achieved by making the marks smaller, but will require an additional factor in light intensity, which can be achieved by speeding up the CRT raster. Faster access also requires * greater requirements of other components, but this can be obtained by further development of other existing techniques.

4. The inter-relations of the design factors may be quite complicated but the problems can be worked out. Also the components presently being used in the system are not the only ones that can be used. Magnetic tapes, punched cards, magnetic cores are only a few of the other components that may be integrated in the system.

Combination of Photoscopic and Magnetic Storage

1. Photographic recording of digital information cannot easily be erased. If changes occur frequently, the photography of intermittently received new information might be rather cumbersome. In these circumstances, it would be extremely convenient to have an associated store into which new information could

be quickly inserted. A magnetic storage unit with a capacity of a few percent of the photo of the photoscopic store would, as a rule, accommodate several days accumulations of new and changed information. When full, its contents would be transferred to a photoscopic disc. Searching of the information could be made in both the photoscopic and the erasable store simultaneously.

Applications

1. The Photoscopic Information Storage Unit can be used in a number of different ways, depending on the nature of the information, the degree of organization, and the rate of change and the rate of growth of the information to be searched*

Automatic Reference Library

1. A difficult problem in documentation or human communication is to assimilate and correlate the vast quantities of information collected. Development of equipment or devices which could perform this function should have a tremendous payoff in research and development.

2. With a few modifications the Mark *I* Translator could be a valuable aid to a research analyst. Large quantities of information can be searched at a very high rate (160,000 characters/sec. or 20,000 words /sec.) for meaningful correlations of facts. Since the photoscopic memory can be searched in a serial manner at a million bits per second, a great number of cross references can be used without the difficulties of multiple filing of all of the entries. In short, the high reading rate is exchanged for collation*

Conclusion

The Ideal Automatic Language Translator

1. The ideal system for performing automatic language translation could have a printed character recognizer which would automatically scan the Russian text, recognize the printed characters of the Russian language and transform these characters into a suitable binary code. This would be done at speeds capable of keeping pace with the output equipment of the automatic language translator. Of course, the output equipment would be selected to fit the particular application. Automatic output equipment can display information at a rate over 10,000 characters per second or print out an original and *6* carbon copies at 1000 lines per minute. Devices of this general type have been developed.

2. The ideal automatic language translator will be a device which achieves the same results as the successful human translator but in a shorter period of time. The automatic translator will be a group of devices which may be mechanical, electromechanical, magnetic, optic or electronic in structure. The language translation system will consist of specialized graphic-semantic information data handling equipment. This equipment is similar in kind to other data-processing devices since the elements are input equipment, storage equipment, data-processing equipment, and output equipment. The equipment is dissimilar to most scientific business data handling computers since language translation machines do not need arithmetic units but are non-mathematical information data handling machines capable of processing semantic information.

References

References on Machine Translation

1. USAF Automatic Language Translator, Mark I - Dr. Gilbert W. King*, IBM, Ossing, N.Y., is the original designer of the USAF Language Translator, Mark *I* and has personally contributed much of his time to further automatic language translation in U.S.A. *Formerly with International Telemeter Corp.

2. Booth, Donald A., and William N. Locke, Machine Translation of Languages, (book), Technology Press of the Massachusetts Institute of Technology and John Wiley and Sons, Inc., New York, 1955.

3. Mechanical Translation. Vols 1-3, 1954-1956, published at the Massachusetts Institute of Technology.

4. King, Gilbert W., G. W. Brown, and L.N. Ridenour, "Photographic Techniques for Information Storage", Proceedings of the I.R.E., Vol. 41, No. 10, October 1953, Page 134.

5. Technical Reports, No. *l-24,* Computer Set, General Information Data, AN/GSQ-l6(XW-l), submitted by International Telemeter Corp. to Rome Air Development Center, Contract AF30(602)-1566, November 1957.

6. Monthly Progress Reports, No. l-l4, Machine Translation Project, submitted by University of Washington to Rome Air Development Center, Contract AF30(602)-1566, August 1957.

7. Appendix A, "Machine Translation of Languages" submitted by the Ramo-Wooldridge Corporation to Rome Air Development Center, Contract AF30(635)-2867, March 1957.