THE PHOTOSCOPIC LANGUAGE TRANSLATOR

Neil Macdonald
Assistant Editor, Computers and Automation

I. General

In May, at the Mohansic Research Laboratory in Yorktown Heights, N.Y., the International Business Machines Corporation in cooperation with the U.S. Air Force, demonstrated publicly for the first time a machine that has been translating the Russian newspaper, *Pravda,* into rough but meaningful English since June 1959.

The Russian translation program at the laboratory has been carried out mostly under contract to the Air Force's Air Research and Development Command, but is also being supported by IBM. A Russian dictionary of some 55,000 word stems has been compiled so far. With the word endings also listed, this corresponds to about a half million words as they appear in text. Ultimately, the Russian dictionary may contain 400,000 stems.

The machine is capable of translating at the rate of 30 words a second. Currently, however, translation is restricted by the speed of the punched paper tape input and the electric typewriter output. The punched paper tape input is produced currently by a typist who types the Russian characters of the text copying them; the machine she is typing on produces in paper tape the binary punched code for each character.

The heart of the electronic translating system, however, is the rotating glass disc called the "photoscopic disc memory." shown on the front cover of this issue of Computers and Automation. About 300,00 coded words are arranged in about 700 concentric tracks, each formed of a sequence of black and white squares, .00033 inches on a side. These are the binary coded forms for the Russian characters and punctuation marks. A sharply focused light beam quickly scans over these tracks, chooses the right track, and matches the Russian input word to the Russian dictionary word on that track. The disc rotates about 23 times a second. The matching is done in less than 1/20 of a second. Corresponding English words are then printed out immediately by an electric typewriter. If at any time the machine finds a word not in its vocabulary, it prints this word in red for later addition to the disc memory. It also prints in red any proper names or nouns which must be "transliterated" or changed from the Russian alphabet to ours.

Next Stages

IBM is also pursuing its own research in automatic reading of printed matter and is supporting a development program at the Baird-Atomic Corporation, Cambridge, Mass. By the end of this year, it is expected that electronic equipment will automatically read and recognize printed material and feed it to the computer at a rate of 40 words a second.

The grammatical capacity of the electronic translator is still well below college level, but a lexical buffer and sophisticated word analyzer now under development will improve this by next year.

The analyzer contains specially-designed logic circuits for analysis of sentence structure and is expected to provide smoother translation from Russian to English than is possible with the present machine.

A new, solid-state, transistorized version of the language translator, the Mark II, is being manufactured at the IBM Federal Systems Division plant in Kingston, New York. This will be faster, extremely reliable, and much more compact than the Mark I model now operating. It is believed that this machine alone will handle all of the U.S. government's presently known translation needs.

Background

Today, less than one percent of the world's foreign technical literature is being translated into English. Only a small portion of this gets to those who need it.

In the Soviet Union, over 2,600 full-time people at the All-Union Institute for Scientific and Technical Translation, plus 26,000 part-time scientist translators publish yearly one-half million abstracts of translated books and articles. The advantage of this in terms of scientific and technical progress has been demonstrated repeatedly. The United States is unlikely to catch up using conventional translation methods. It is for this reason that IBM and the Air Force have felt the need for automatic language translation.

While the Air Force's main interest in machine translation is in the area of Russian technical information, it is of course true that other areas in industry, science and government already desperately need new ways of storing and retrieving the constantly changing and increasing masses of information. Automatic language translation is one of the first steps in organizing, finding, and disseminating the latest information in fields as diverse as cancer research, physics, chemistry, economics, and space technology. Russia alone publishes several billion words a year in such areas.

Less than one tenth of one percent is currently translated — at a present cost of about eight to ten cents a word to the United States Government. The electronic translator can cut costs to a very small fraction of this. But even more important are the many important Russian papers which could be translated, but which are now left untouched due to lack of time and money.

Early work done by Dr. Gilbert W. King in this field was carried out at the International Telemeter Corporation. In addition, the Air Force contractors have included Cambridge University (England), Harvard University, University of Indiana, University of Milan (Italy), New York University, Syracuse University and the University of Washington, as well as Baird-Atomic Corporation, Intelligent Machines Research Corp. (now part of Farrington Mfg. Co.), Thompson Ramo-Wooldridge, and International Business Machines Corporation.

Machine Translation from French to English

At the same time, IBM also reported progress in a program of machine translation of French undertaken by the company on its own. In this program, a dictionary of about 23,000 French words and their English equivalents is already being used to produce rough translations of mathematical papers. This machine translator's vocabulary will be increased shortly to permit translation of non-mathematical subjects as well.

The present quality of the machine translation, which is similar to the present quality of the machine translation from Russian to English, is indicated in the following example, which consists of A, a passage in French, and B, the translation in English.

A

DEFINITION DE LA LOGIQUE MATHEMATIQUE. La logique algébrique qui est le sujet de ce cours, est conçue ici comme la partie la plus élémentaire de la logique mathématique. Plus tard nous préciserons ce que nous entendons signifier par le mot "algébrique." Mais il faut indiquer tout de suite en quoi consiste la logique mathématique dont la logique algébrique constitue la première partie.

Dans cette intention rappelons que le mot "logique" a trois sens différents dans presque toutes les langues.

B

DEFINITION (OF) THE MATHEMATICAL LOGIC

The algebraic logic which is the subject of this course/s is conceived here as the part the most elementary (of) the mathematical logic. Later we/us will specify what we/us hear/mean signify by the word "algebraic." But one needs indicate immediately in what consists the mathematical logic whose algebraic logic constitutes the first part.

In this intention recall that the word "logic" has three different sense/s in almost all the languages.

II. Some Technical Details

The photoscopic disc, which acts as an automatic dictionary, is read by means of a cathode-ray tube light source, a moving lens, a photo multiplier tube, and some electronic circuitry (the digit detector). In addition to the disc equipment, key portions of the system include the input Flexowriter, which is used for typing the input Russian text, the input register which holds input text until the disc is searched for the proper dictionary entry, and the output Flexowriter which prints the translation.

Input

The input Russian characters are each coded in the form of holes in an input tape. After the tape passes through the tape reader, the information becomes coded into "ones" and "zeros," six per character, and is placed in the input register.

Comparison

These characters are then compared with the information being read out of the dictionary in order to determine the proper direction to move the lens and cathode-ray tube beam. The light beam continues to step across tracks, reading a small portion of each, until the comparator indicates that it has gone too far. The light beam is then brought to rest and the disc rotation (23 r.p.s.) allows the reading of every entry on a particular track. This corresponds approximately with our reading the entries on a page of a printed dictionary from the bottom up in order to get the longest possible match (for example, "time constant" before "time"). When a proper match to a Russian semantic unit has been found, the corresponding English meaning is read out through the high-speed register to the output Flexowriter. At the same time, logical circuitry indicated by the "distributor" has kept an account of the number of input characters for which a match has been found. This allows the input characters which have just been translated to be discarded and fresh input characters to be shifted into the input register.

Output

An output buffer is the gathering place for all output characters prior to printing out. It holds one character at a time, receiving it from various sources depending upon the nature of the contents. When no translation whatsoever is required, such as for Roman characters, punctuation, and numerals, the characters come directly from the input register. In the case of Russian word inputs, English meanings come to the output buffer from the disc memory equipment. In the case of input Russian which must be transliterated into Roman characters, such as proper names, the direct transfer from the input register is blocked so that control circuits can allow the transliteration stuffing equipment to make the proper changes. When an input Russian word should be translated but cannot be found in the automatic dictionary, it is also transliterated. In both cases of transliteration, the output typewriter ribbon color is shifted to red. This feature allows the user to notice at a glance what additions are required in the next edition of the dictionary.

Light Source

A cathode-ray tube is used as a light source because an electron beam can be moved faster than any other source. When it is necessary to move the light from one disc track to another, the change is made extremely rapidly by means of the deflection current. The lens motor, with its higher inertia, moves more slowly and allows the cathode-ray tube electron beam to return to the center of the tube face. Thus the cathode-ray tube beam makes possible the low access time (35 milliseconds average) while the lens motor prevents the electron beam from going too far toward the side of the tube.

Speed

The speed of this translating system is sufficient to translate Russian technical literature at an average rate approximately equal to the rate at which it is produced (30 words per second at present), except that the relatively slow speed of the input and output Flexowriters restricts the speed of the system to a substantially lower rate. In the future it is expected that the input speed limitation will be removed by the use of automatic page scanners and character sensing. The output speed limitation can be eliminated by means of high speed printers and by multiplexing several output units.

Memory

The ten-inch glass disc contains 30 million bits of information on 700 tracks in an annulus 0.36 inches wide. The width of the annulus is kept as small as possible for the sake of rapid access. The coded lexicon entries are represented by clear and black squares approximately one-third of a mil (0.00033 inches) on a side. These marks interrupt the light beam from the cathode-ray tube on its path to the photomultiplier and, therefore, differentiate a zero from a one. Since these marks have been kept small for the sake of access time, the light source on the face of the cathode-ray tube has also been kept small. It is, however, quite intense (1600 candle power). In order to prevent the electron beam from burning the phosphor on the face of the tube, as will happen in any cathode-ray tube when its beam is kept in the neighborhood of a single spot, the beam, in addition to the motion required for stepping from track to track, is made to trace a circle about two inches in diameter. This motion is compensated by a vibration in the lens so that positioning on the tracks of the disc is not affected.

Addresses

Lexicon entries on the disc are laid out in such a way that the Russian words and idioms themselves make up the address of the entry. Each character in the Russian word has a certain binary code which can be interpreted as a weight. Cyrillic "e" has the lowest weight and Cyrillic "B" the highest. Each coded Russian word, therefore, looks like a long binary number. The layout on the disc is in numerical order, which is different from alphabetical order but has a kind of similarity to alphabetical order. When the disc is being scanned track by track, each bit (one or zero) in the Russian word is compared with the corresponding bit in the input register (which here is acting like a memory address register).

This comparison is continued until disagreement is found. At this time, a "zero" on the disc and a "one" in the input register means "go ahead" to the next track. The inverse combination means "go back." Only when the correct entry has been passed is a particular track scanned exhaustively at a rate of one million bits per second. The exact match has been found when each one and each zero in the Russian dictionary word matches up exactly with each one and zero in the input register until the symbol signifying the end of the Russian word in the dictionary entry is reached.

The search electronics could readily be reduced in size by converting the high speed vacuum tube circuits to transistor circuits.

The design of the dictionary has utilized photographic techniques because photographic emulsions are the densest storage media know today. Although this storage is permanent, a great deal of work has gone into the design of equipment which can prepare new discs rapidly, thus allowing frequent updating of the list of entries in the dictionary.

Preparation of the Memorized Dictionary

The system for preparation of the disc memory begins with a document which contains entries to be placed in the dictionary. These entries are keypunched into cards. Once the cards have been prepared, an IBM 704 computer merges new cards with old by straightforward sorting procedures, placing the complete list of entries on magnetic tape. The 704 sorting program also will place any random mixture of dictionary entries into proper numerical order. Especially designed equipment allows the magnetic tape to serve as input to a film-making machine which photographically codes entries on rolls of 70-mm film. The use of film as a preliminary step prior to making the disc substantially reduced the initial engineering problems (such as timing) and allowed faster progress in building a working system

Since film-making is a photographic process (a cathode-ray tube light source is turned on and off in accordance with coded binary information arriving from the magnetic tape while the film is passed through the machine), photographic processing must follow the exposure. Although a standard Air Force photographic processor is used at present, some improvement appears desirable for future systems. Toward this end, research is developing a special film processor which makes the processing automatic.

Information on the film is recorded on the disc by means of a Disc-Making Unit. Here again a cathode-ray tube light source is used, along with a mirror and lens system for passing light through the film to the emulsion on the disc. A servo system is utilized for accurate focussing. The disc is turned at one revolution per minute during exposure; special precautions have been taken in the design to keep the speed and timing within close tolerances. The size of the photographic spots of information is reduced by a factor of about 55 during this process.

The photographic processing of the disc after the disc-making machine records the information is extremely sensitive to foreign particles in the developing and fixing fluids, due to the extremely small size of the information bits (0.00033 inches square). Commercially available solutions are completely useless until they are finely filtered. To overcome this problem, a special disc processor was designed. This device minimizes the quantity of fluid used and then filters immediately before the fluid touches the emulsion surface.

As a result of the great care taken in the design of the disc-loading system, new dictionaries can be prepared on approximately a daily basis if desired. This speed meets the requirements of language translation by a wide margin. However, even if more rapid updating of the storage is required, a small erasable store such as a drum could easily be added. In this system, the input address would initiate simultaneous searches in the large capacity photostore and in the small capacity erasable store. In most cases information would be found in the photostore only. Whenever a match would be found in *both* memories, the "logical choice" equipment would dictate that the erasable store be read out, since its information would be the most recent.