[From: *Computers and Automation*, vol.2, no.5, 1953]

MACHINE TRANSLATION

by Yehoshua Bar-Hillel
Research Laboratory of Electronics, Mass. Institute of Technology, Cambridge, Mass.

(The work reported in this article has been supported in part by the Signal Corps, the Air Materiel Command, and the Office of Naval Research, and in part by the Rockefeller Foundation.)

More than a year ago, I wrote a paper entitled "The Present State of Research on Mechanical Translation", which was published in "American Documentation" (see reference 8 below). Few engineers perhaps are likely to consult the journal in which it was published, and additional advances have been made during 1952 and the first months of 1953, both in theory and in organization. Therefore it may be worthwhile to present here a summary of my earlier paper and to indicate some of these advances.

Among the noncomputational applications of high-speed digital computers, facilitation of translation from one language into another was considered relatively early. Indeed, a number of the operations that are performed during a complete translation process are routine operations, or are at least replaceable without loss by routine operations. Besides, there is a severe shortage of bilingual persons who can translate scientific material accurately and who can scan at high-speed the printed output of actual or potential enemies. This shortage has served as a potent incentive for research into the possibilities of total or partial replacement of human beings by automata in translation of languages.

Let me describe here somewhat dogmatically, for lack of space, the present outlook. Fully automatic high-accuracy translation seems out of the question in the near future. If this aim were achieved, it would require either storage capacities of trillions of bits of information, or the invention of programming techniques that could greatly increase the machine's efficiency by learning, or some combination of these two factors. At any rate, we would need something that would resemble a human being in versatility much more than automata will be able to do in the reasonably near future. Therefore, either the high accuracy or the complete automatic character of the translation process must be sacrificed.

Whenever high accuracy is essential, as is the case generally in translation of scientific material — and I mean here translation proper and not merely scanning for worthwhileness of careful translation —, only man-machine partnerships are at present in view. These may still be of quite diverse characters. One combination is known as "machine plus post-editor", where the post-editor takes the machine's first approximation and knits it together. A second combination is known as "pre-editor plus machine"; here the pre-editor rearranges the foreign text into a natural form for the language into which it is to be translated. A third combination is "machine plus bilingual editor" (whose time of employment will, of course, be highly restricted so as not to trivialize this kind of partnership). These combinations have already been explored to some extent.

Each of these partnerships offers its special advantages and hardships. It seems that a post-editor would be at his best in eliminating semantic ambiguities; a pre-editor would conceivably excel in rephrasing to avoid grammatical ambiguities and in indicating "idiomatic" expressions; and a bilingual editor might be profitably employed to deal with stubborn "remainders".

The partnership of machine plus post-editor seems to me of special theoretical interest. It appears, incidentally, that this might also be the most helpful practical combination. For example, there is no shortage of expert English-speaking chemists, but there are not enough of them who also know, say, Russian to a sufficiently high degree.

What, then, could a machine do to aid the post-editor in producing a satisfactory counterpart of, for example, a certain German paper? (I assume that we shall soon have mechanisms that will be able to "read" printed material.) The device that comes immediately to mind would then be a mechanical dictionary, that is, an apparatus which correlates to the coded version of each German word the coded version of one or more English words or phrases, or combinations of such with what we might call operators. A word, in this context, means anything that appears between spaces, or between a space and a period, and the like, that is, something identifiable by its shape alone.

The decoded form of an entry in such a dictionary might look like this: "lieben — (1) love, (2) to love, (3) dear"; or, alternatively, "lieben — (1) love (inf.),(2) love (pres., plur., 1st or 3rd person), (3) dear (plural), (4) dear (sing., gen. or dat. or acc.-masc.)" with the "operator" in parentheses.

This type of system would require a storage of between one and two million entries and their correlates, hence of hundreds of millions of bits under ordinary alphabetical coding. Since the access-time has to be, for practical reasons, of the order of tenths of seconds at most, the preparation of a mechanical memory adequate for this task presents a serious, though certainly not insurmountable, engineering problem. I shall not discuss here how coding skill and intelligent organization of the dictionary could reduce both storage capacity and access-time (but see reference 7).

Experimentation carried out by different groups (ref. 5a, 6, 7) seems to indicate that, for translation from Russian into English, from German into English, and from some other Indo-European languages into English, the output of the mechanical dictionary, at best arranged in columnar form, is often sufficiently intelligible to the expert post-editor so that he can write down almost immediately a unique (up to the point of synonyms) translation of the unknown original. The following is an example of a simplified hypothetical output of a German-English mechanical dictionary, consisting of what the machine would conceivably present to the post-editor as its first approximation in English to a given German sentence. The reader would do well to scan the alternatives and then write down for himself what he prefers as the translation of the unknown German sentence, before he consults the German original stated at the end of this article.

```
the               answer      on                   this            question
which (rel.)     reply       (any preposition)    this one         problem
who (rel.)                                         the latter      demand
                                                                   inquiry


hang (pres., 3rd, sing.)           both                  at the
hang (pres., 2nd, plur. )          as well              (any preposition) the
depend (pres., 3rd, sing.)
depend (pres., 2nd, plur.)

optical microscope       as           also              at the
                         since                          any preposition) the
                         when
                         than
```

electron microscope	from	three	different
	(any preposition)		differing
			unlike
			various
			deceased

property (plur.)	of the	microscope	from
quality (plur.)			upon
attribute (plur.)			(preposition or
character (plur.)			separable prefix)
nature (plur.)			
condition (plur.)			

The reader will have noticed, to his annoyance, the occurrence of one of the most disturbing features of German construction, the separable prefix. He will also have noticed the special difficulties involved in the translation of prepositions. I anticipate, nevertheless, that the translation with which most readers will wind up will not be far off the point, in spite of the fact that they have to choose between some 50,000 combinations (disregarding the choice of the prepositions which, if taken into account, would have increased this number to many millions). However, the given sentence is only of average complexity or less. Translation in this fashion of complex sixty-or seventy-word sentences—which are not too infrequent in German scientific writing—would have presented much greater troubles. It might turn out that the load on the post-editor would be too great for any practical purposes.

Could not something more perhaps be done by the machine? Could it not eliminate grammatical ambiguities—such as in the first column, where only the definite article is acceptable, or in the sixth column, where only the 3rd person fits? Could it not rearrange the words into some standard English word-order—-such as the rearrangement of "hängt... ab" into "abhängt"? I think this is definitely possible, but only after much linguistic spadework of a type to which linguists have not been accustomed so far. Elsewhere (ref. 9) I have attempted to give the outlines of such a new approach to linguistic analysis. At another place (ref. 8), I exhibited an output of a hypothetical "mechanical analyzer plus rearranger plus dictionary" which was tested on a small scale and shown to be fairly satisfactory. At still another place (ref. 10), I discussed other aspects of this problem, as well as some methods of dealing mechanically with idioms.

Lack of space prevents us from going into further details here of research done so far on machine translation. The attached annotated bibliography should, however, enable the interested reader to get more of the necessary information.

Let me conclude my outline by mentioning the Conference on Machine Translation. This conference met in June 1952 at Massachusetts Institute of Technology under a grant from the Rockefeller Foundation. Almost everyone who is actively engaged in research on machine translation participated at this conference, and also many others whose interest until then was more academic. This conference provided the participants with much stimulation and actually induced some outsiders to take up research in this field. It was the consensus of all participants that machine translation shows definite possibilities. Various projects are now under way but none, to my knowledge, is being undertaken on a scale that would ensure rapid progress. However, I am convinced that within a decade, at most, substantial achievements toward machine translation will have been made. It is tempting to speculate about the sociological feedbacks of this development, but I prefer to remain, for this account, within the limits of factual sobriety.

Here is the German sentence whose hypothetical correlates through a mechanized dictionary were presented above:

"Die Antwort auf diese Frage hängt sowohl beim Lichtmikroskop als auch beim Elektronenmikroskop von drei verschiedenen Eigenschaften des Mikroskops ab."

Bibliography

1. Warren Weaver, Translation, mimeographed, 12 pages, July 15, 1949. — This memorandum contains an account of the early history of machine translation and comprises, among other things, an interesting exchange of letters between Weaver and Norbert Wiener in 1948; Wiener was on the skeptical side. Weaver also mentions a memorandum by A. D. Booth, now Director of the Computation Laboratory in Birkbeck College, University of London, dated February 12, 1948, in which translation with the help of a mechanized dictionary is considered. Weaver, however, was apparently the first to consider machine intervention going beyond a mechanized dictionary. Though Weaver's contribution was admittedly of a rather speculative nature, it gave, no doubt, the major impetus for subsequent research on machine translation in the United States.

2. Abraham Kaplan, An Experimental Study of Ambiguity and Context, the RAND Corporation, P-187, 18 pages, Santa Monica, November 30, 1950. — This monograph is an outcome of a study undertaken at RAND with a view to rapid mass translation. Though of only direct interest to machine translation proper, its contents give a possible theoretical background for the explanation of the speed and efficiency with which a post-editor can select a good translation out of millions of candidates.

3. Victor A. Oswald, Jr., and Stuart L. Fletcher, Jr., Proposals for the Mechanical Resolution of German Syntax Patterns, Modern Language Forum, 36: 1-24, 1951. — This paper shows how certain syntactically ambiguous word sequences can be mechanically uniquely resolved. Though admittedly only a first attempt, it contains many nuclei for a more systematic attack.

4. Erwin Reifler, Studies in Mechanical Translation, mimeographed, 1951-53. — The author embarked, under the influence of Weaver's memorandum, upon many ingenious studies dedicated to machine translation. Thus far he has published eight studies in mimeographed form and has many more in preparation. He is also preparing a revised version of these studies for publication in printed form.

His ideas are too manifold to be summarized here. Let it be said only that in his first studies he dealt mainly with the problems of pre-editor plus machine partnership, but he has now shifted his attention to the machine plus post-editor combination with a minimization of the latter's participation in view. Insufficient acquaintance with the working of digital computers reduces somewhat, though by no means nullifies, the impact of his studies on immediate applications.

5. Conference on Mechanical Translation, M.I.T., June 17 - June 20, 1952. The Proceedings of this conference are due to appear eventually. Only a few of the papers submitted in mimeographed form are summarized and evaluated here.

a. R. H. Richens and A. D. Booth, Some Methods of Mechanized Translation, 31 pages — The authors are the main adherents of the thesis that a mechanized dictionary, supplemented by a limited amount of mechanized grammatical analysis, mainly suffix analysis, is in general a sufficient basis for subsequent post-editing. The thesis is

illustrated by translation specimens involving some twenty languages. Schedules for punch-card and electronic machines are outlined.

In my opinion, the general redundancy of syntactical analysis prior to word-by-word translation has not been sufficiently substantiated by the authors.

b. Victor A. Oswald, Word-by-Word Translation, 8 pages. — Oswald, in direct opposition to Richens-Booth, tries to show that word-by-word translation will not, in general, yield intelligible outputs. It seems, however, that this evaluation is too pessimistic, just as the one given by Richens-Booth was too optimistic. Obviously, much more extensive experimentation is required to settle this question.

c. Victor A. Oswald, Microsemantics, 10 pages. — As a partial aid in reducing the overload of correlates presented by straightforward word-by-word translation, Oswald proposes the construction of special glossaries, since many words, which are highly ambiguous in general, lose much of their ambiguity when it is known that they are used in a study belonging to a certain restricted field.

d. Charles S. Dodd, Model English for Mechanical Translation, 9 pages. — The author foresees the impact of regularizing, prior to translation, any or both of the languages involved in mechanical translation. He expounds in detail a method of "modelizing" English which would highly increase its grammatical regularity and thereby reduce the machine effort of finding out the grammatical structure of a given sentence.

e. Other participants dealt with Operational Syntax, Treatment of Idioms, Mechanical Translation of Printed and Spoken Material, Frequency Problems in Mechanical Translation, Teaching of Foreign Languages, Basic Machine Operations in Mechanical Translation, Problems of Storage and Cost.

6. James W. Perry, Memoranda on Mechanical Translation, mimeographed, 1952-53. — In the five memoranda prepared so far, the author discusses mainly the various possible designs of mechanical dictionaries and the impact of grammatical indicators on the intelligibility of the machine output. One of his results states that omission of some grammatical information does not seriously reduce the intelligibility of a word-by-word translation from Russian into English, in corroboration of the Richens-Booth thesis.

7. A. G. Oettinger, A Study for the Design of an Automatic Dictionary, Progress Report No. 26, Computation Laboratory, Harvard University, February 10, 1953. — The author, who is working on a Ph.D. thesis on machine translation, independently discusses problems similar to those treated by Perry but he gives special attention to coding.

8. Y. Bar-Hillel, The Present State of Research on Mechanical Translation, American Documentation 2:229-237, 1951 (appeared 1953). — This report was prepared in January 1952 and widely circulated in mimeographed form. It contains the first synopsis of the various approaches to machine translation.

9. Y. Bar-Hillel, A Quasi-Arithmetical Notation for Syntactic Description, Language, 29:47-58, 1953. — A method is described whereby syntactical analysis of any given sentence is possible if a complete list of the syntactic categories to which all words of the given language may belong is prepared.

10. Y. Bar-Hillel, Some Linguistic Problems Connected with Machine Translation, Philosophy of Science, 1953. — The problems of the need for an operational syntax, of the treatment of idioms in machine translation, of the intertranslatability of all natural languages, and of the existence of a universal scheme of syntactic categories are discussed.

11. Kenneth E. Harper, The Mechanical Translation of Russian: A Preliminary Report, University of California, Los Angeles, 26 pages, 1953. — The author shows convincingly that machine translation of scientific Russian, on the basis of an idioglossary and suffix analysis, is satisfactory and immediately feasible. "It appears that a Russian vocabulary of some 3000 words would be adequate for the translation of mathematical papers."

12. Victor A. Oswald, Jr., and Richard H. Lawson, An Idioglossary for Mechanical Translation, University of California, Los Angeles, 16 pages, 1953. — The authors prove experimentally that an idioglossary of less than 5000 entries; suffix analysis, and a restricted syntactic analysis form a sufficient background for effective mechanical translation of German papers dealing with brain surgery. The statement, however, that "the major linguistic problems of pragmatic mechanical translation can now be regarded as solved" is not completely borne out by the presented facts.

In this bibliography only material prepared with machine translation as its main object is considered. Some other relevant publications are mentioned in the bibliography of reference 8.