Machine Translation*

By Erwin Reifler, Dr. Rer. Pol., *Professor of Chinese*
Director, Chinese-English Machine Translation Project
University of Washington, Seattle, Washington

Introduction

THE possibility of using electronic calculators for the mechanization of the translation process was first seriously considered early in 1945. It occurred to Dr. Warren Weaver, then a director of the Rockefeller Foundation, that it might be possible to make use of, or to design, such machines for the high-speed mass translation of scientific publications. On July 15, 1949, Dr. Weaver wrote a memorandum, entitled *Translation*, which he sent to some two hundred scholars working in various fields. In this memorandum he mentioned his attempts to interest experts in cybernetics in his idea, and, fully aware of the complexity of the linguistic problems involved, he outlined several lines of attack in which due consideration is given to the phenomenon of language universals, to the problem of multiple meaning, and to the pinpointing function of environment and context.

Dr. Weaver sent me a copy of his memorandum because he referred in it to a paper which I read in April 1947, before the American Philosophical Society in Philadelphia on "The Chinese Language in the Light of Comparative Semantics," in which I gave numerous examples demonstrating agreements between unrelated languages which are not due to borrowing or cross-fertilization. (An abstract of this paper was published in *Science* 107: 586, June 4, 1948.) Dr. Weaver used one of my comparative semantic examples in support of one of the four lines of attack he suggests for the solution of the automation of the translation process.

Let me give you some examples for such extraordinary agreements between unrelated languages. A striking example is Chinese *t'ung* (童), which means "child" as well as "pupil of the eye" (in the latter sense today written 瞳 —that is augmented by 目 which means "eye"). Many languages share this phenomenon of the association of the two, apparently incompatible, notions of "child" and "pupil of the eye" in one and the same word. English "*pupil*", which also means "school child," is itself derived from Latin *pupilla* meaning "little girl" (*pupillus* means

*Presented at the Sixtieth Annual Meeting of the Medical Library Association, Seattle, Washington, May 7–12, 1961.

"little boy") as well as also "pupil of the eye." The explanation for this phenomenon is the fact that whenever we look into somebody's eyes, we see there a small mirror image of ourselves.

Another example is Chinese *hsiang* (鄉) which means "opposite, against" as well as "country" (as opposed to "town"). This is paralleled in English, French, and German: compare English *country*, which is derived from French and of the same origin as English *counter-* in *counter-attack* and *contra-* in *contradict*, and German *Gegend* which means "countryside" and is derived from *gegen*, meaning "opposite, against."

A third example is Chinese *ts'ung* (葱), meaning "onion" and *tsung* (總), meaning "unite, union." The two Chinese words are cognates. The association of the two apparently incompatible meanings of "onion" and "union" is paralleled by the English words *onion* and *union*, both derived from the one Latin word *unio* which already had these two meanings. These parallel developments in the evolution of meanings of words of unrelated languages are independent in the sense that they are linguistic coincidences, although they betray the workings of a common human logic. This is an important point in the development of machine translation and that is the reason Dr. Weaver referred to it in his memorandum.

This memorandum soon began to produce results. At Massachusetts Institute of Technology, at the University of California, Los Angeles, at the University of Washington in Seattle, linguists and engineers formed teams for the joint investigation of this new field. In June of 1952 the First International Conference on Machine Translation was held at MIT. This was followed by a number of research grants from the Rockefeller Foundation, two of these, in 1952 and 1953, given to me. In January, 1954, the World Headquarters of IBM of New York announced through the newspapers, radio, and television all over the world that a 701-type calculator had been applied to translation work and had actually translated a number of specimen Russian sentences into English. Thus the world at large learned for the first time of the coming miracle of machine translation. The preparatory studies for this IBM experiment had been done at the Institute of Languages and Linguistics of Georgetown University in Washington, D. C. In the same year the first issue of the journal for MT development, called *Mechanical Translation*, published at MIT, made its appearance.

Of great importance for the further spread of MT research was the joint publication in 1955 by the Technology Press of the MIT and by John Wiley and Sons of the first volume on the subject. This book, entitled *Machine Translation of Languages*, is a collection of essays by 14 pioneers who took part in the International Conference at MIT in 1952. From

the appearance of this volume dates MT research in Europe, especially in the Soviet Union, where theoretical and experimental work was initiated in 1955.

Since that time, machine translation research reports have been on the program of the meetings of many academic societies, and an ever-increasing number of national and international conferences is dedicated to this new field of applied linguistics and engineering, for example the International Conference on Machine Translation of Languages and Applied Language Analysis in Teddington, Middlesex, England, September 5–8, 1961. At present machine translation research is being carried on in many universities, government agencies, and private corporations in this country and abroad.

At present we have difficulty coping with a growing avalanche of publications on MT from many countries, and millions of dollars have already been spent and are being spent by governments, foundations, educational institutions and industry in different parts of the world. Examples are the Russian-English and the Chinese-English Machine Translation Projects of the University of Washington, both of which I have been directing. The first has been sponsored by the United States Air Force which, since May of 1956, has already given this University $292,000 for this purpose. The second has been supported by the National Science Foundation with a grant of $53,700 for one year.

In 1958 we published a comprehensive 660-page report on our project. It is primarily concerned with the first phase of our research, the lexicographical phase and the engineering problems involved. It does, however, also deal with initial research results in the automatic resolution of grammatical and nongrammatical ambiguities.

In October, 1960, our University Press published our second comprehensive report, totaling 504 pages. It deals with the structural-analytic phase of our research and our further experiments in the automatic resolution of ambiguities. Our report on the Chinese-English project is about to be published.

The best outline of MT research in the world at large is a recent book by Émile Delavenay, Chief of Documentation of UNESCO, entitled *La Machine à Traduire*. Praeger in New York has brought out an English translation under the title *An Introduction to Machine Translation*.

The Linguistic Problems

Machine translation development today is something that first has to be solved on the linguistic level. Once the linguistic problems have been solved, the engineers will know how to solve the engineering problems involved.

The paramount linguistic problems are those presented by the *conventional symbolization* of languages and those of *multiple meaning* in terms of at least two languages. This problem of multiple meaning in terms of two languages I call *source-target semantics.* The latter problem has to be considered on two levels: the *grammatical* and the *nongrammatical.*

The Problem of the Conventional Graphic Form. The first problem is that of the conventional symbolizations of languages. There can be no doubt that both their spoken and their written forms fall within the sphere of interest of mechanical translation as chief means for a mechanical correlation. But the conventional phonic symbolization of free forms is in a number of important languages often less distinctive than their corresponding written representation. Striking examples are English "to, too, two." Homophony plays an even greater role in languages like Chinese and Japanese. In such languages the "historical" form of writing is symbolico-semantically much more distinctive. There are, of course, also many cases of the converse phenomenon of different pronunciations of words with identical spellings. Examples are English "the bow" and "to bow," "the wind" and "to wind," "the sow" and "to sow," "the tear" and "to tear." However, as a result of research I conducted during the summer of 1953 under a grant from the Rockefeller Foundation I can state that there is a very simple mechanical solution for, at least, a large number of these ambiguous cases. It is, on the other hand, a well known fact that the graphic form of language is generally much more explicit and leaves much less to be inferred from situational criteria than its spoken form. Thus it will mostly present MT with a much less formidable problem. Consequently, the MT linguist will—at least at this initial stage of mechanical translation development—mostly study language in its conventional *graphic* form where he is not concerned with *homophones* but with *homographs.*

The Problems of Multiple Grammatical Meaning and Model-Target Languages. It was clear from the outset that the problem of multiple grammatical meaning as exemplified by the English "considered," which is either the past tense of a verb as, for example, in "he considered," or a qualifying adjective as in "this is his considered opinion," presented much fewer and smaller difficulties than the problem of multiple nongrammatical meaning as exemplified by English "date," denoting a time, and "date," denoting a fruit.

Another interesting problem to be mentioned is the mechanical correlation of the grammatical meaning of words of the source and the target language. Many languages are characterized by certain so-called morphological irregularities. Examples are English "boot" and "boots,"

but not "foot" and "foots" nor "boot" and "beet." Other examples are "to link" and "linked," but not "think" and "thinked" nor "think, thank, thunk," as in "sing, sang, sung," nor "to link, lank, lunk," nor "to link, lought, lought," as in "think, thought, thought." If such irregularities could be removed, if the paradigms of such words could be regularized, then the mechanical correlation between two irregular languages, or between one regular and one irregular language could often be made much simpler. I have, however, to stress here the fact that, if we aim at a practical solution of mechanical translation, then we can interfere neither with the conventional written form nor with the grammar of the source language. But on the target side we can, within certain definable limits, plan the form of the target language. We can put a selected vocabularly and a regularized morphology and syntax into the machine and, moreover, within the limitations of intelligibility, adjust the target language to certain peculiarities of each of the source languages.

The Problem of Multiple Nongrammatical Meaning. Dr. Weaver had already suggested in his memorandum that "it does seem likely that some reasonable way could be found of using the micro-context to settle the difficult cases of ambiguity." In one of my papers I outlined, and in subsequent research I further developed, ideas aiming at such an ultimate elimination of the human post-editor. It is clear that a mechanical translation system whose design permits the mechanical determination of intended nongrammatical meaning must be very much more complex. It will necessitate more engineering and require more equipment and mechanical operations. But as long as these requirements do not exceed the boundaries of practicality, I see no reason why such a solution should not be sought. Such a solution would extend the scope of MT beyond its present limitation to scientific publications.

The Elimination of the Human Pre-Editor. The greater mechanical complexity necessitated by the envisaged elimination of the *post*-editor makes economies on other levels of mechanical translation particularly welcome. Such economies presented themselves in my subsequent research. During the summer quarter of 1952, I concentrated on the problem of the elimination of the *pre*-editor. This research was made possible by a grant from the Rockefeller Foundation. It was mainly concerned with the two problems of "unpredictable compounds" and "the mechanical determination of essential grammatical meaning." The results of this research leave no doubt that an MT system can be built which abstracts all essential information from the conventional source text without the necessity of human intervention and that it is possible to substantially reduce the number of lexical items to be coded into the mechanical memory.

The Substantial Reduction in the Size of the Mechanical Memory

The Mechanical Dissection of Complex Forms. The earliest mechanical translation scheme, that developed by Booth and Richens in England, already includes the mechanical dissection of complex forms into stems and endings. This possibility of automatic dissection of, for example, words like *joiner, singer, reader,* into *join, sing, read* and *-er,* permits the omission from the machine memory of the words *joiner, singer, reader,* because these can be automatically identified by the stems *join, sing, read* and the ending *-er.*

The Mechanical Dissection of Known and Unpredictable Compounds. The obvious advantages of such a procedure could, of course, also be made available in the case of compound forms. Three difficulties, however, had to be faced here from the outset. One is the difficulty presented by source language compounds whose target meaning can often not be inferred from the meaning of the target equivalents of their constituents. An English example is "mushroom" which risks being translated as "a room of mush." Another example is "spokesman" for which we might get as translated meaning something like "a man of spokes." English examples of another kind are "teasing" for which we might get a dissection into "tea" and "sing," "taciturn" for which we might get a dissection into "tacit" and "urn," and "season" which could result in a translation meaning "son of the sea." A German example is "Mit/gift," literally "with/poison," but actually meaning "dowry." This difficulty, however, can easily be met by entering all such compounds into the mechanical memory. This leaves only those compounds to be mechanically dissected whose target meaning *can* be inferred from the meaning of the target equivalents of their constituents.

But the other two difficulties seemed, at first, to constitute insoluble problems. The first is the so-called "X" factor as exemplified by German "Dichterinbrunst." Correctly dissected into "Dichter" and "inbrunst," this means "a poet's fervor." But also "Dichterin" and "Brunst" occur in German as free forms, although such a dissection of "Dichterinbrunst" would be morphologically and semantically wrong. It would give us the English translation "a lady poet's sexual desire of a male animal." Such a meaning does not come to the mind of a German who hears or reads that word, although it could, perhaps, be rendered in English by "Sapphic desire." The reason is simply that such a dissection is morphologically wrong. Another German example is "Wachtraum." Here two dissections are permissible both from the morphological and the semantic points of view. These are either "Wach/traum"—that is, a "waking dream" or "day-dream," or "Wacht/raum"—that is, a "guard room."

The second difficulty is that of extemporized—that is, unpredictable—compounds, such as English "holdability."

The results of my research during the summer of 1952 proved beyond any doubt that there is actually a simple mechanical solution even for this problem of the identification of the constituents of *all* compounds which are not "memorized" by the mechanical memory, but whose constituents occur there. My solution permits the complete elimination of the human pre-editor from the identification process of mechanical translation, but also provides the machine memory with a much smaller vocabulary than our comprehensive standard dictionaries. Consequently, the future machine translation industry will be able to save millions of dollars.

Because I am familiar with German, and this language is notorious for its abundance of unpredictable compounds and, moreover, important for machine translation, I first developed this solution for the substantive compounds of the German language and then tested it on other languages. Examples for German extemporized compounds are "Marsuranium-monopolskandal" and "Grieselbaerintelligenzexperiment." My solution is applicable to all languages which have the same problem. As a result, the number of compounds which have to be entered into the mechanical memory can be very much reduced. A few examples will illustrate this. Compounds like English "seashore" (substantive plus substantive), and "cutthroat" (verb plus substantive) need not at all to be coded into the mechanical memory. But, nevertheless, their target meanings can—at least in the case of a large number of target languages—be inferred from the meaning of the target equivalents of their constituents.

These have been examples of well known English compounds. But also extemporized compounds, although not as common as in German, turn up daily in the English language. A striking example, "holdability" I have already mentioned. It occurred in a title on page 11 of the Sunday Magazine of the *Seattle Times* of March 14, 1954. This title read "Nails With More Holdability." Both "hold" and "ability" will, of course, occur as free forms in the mechanical memory. They could in German as the target language, for example, be made to appear as "halt-" and "-barkeit," respectively.

The Elimination of Words of Dual Nationality. My solution of the problem of unpredictable compounds opened the way for the elimination from the mechanical memory of another substantial group of source language forms—the forms of dual nationality and of the compounds containing them. Let me illustrate this with some examples from the French language. My system for the mechanical dissection of compounds is, of course, also applicable to French, but in this language compounds

which are not shared by some other languages are comparatively rare. Most French compounds permitting a morphological and source-target semantic determination through a mechanical synthesis of their constituents occur also in German, Russian, English, and in many other languages—sometimes with minor orthographical and morphological changes. They are "pure (as different from "hybrid") international compounds," and as such no identification via the identification of their constituents is necessary. Neither they nor their constituents need to be coded into the mechanical memory. Examples are "tele/gramme, tele/graphe, tele/graphie, tele/scope, micro/cosme, micro/graphie, micro/metre, micro/scope." With the help of simple matching procedures developed during my research, it is possible to deal successfully with all problems encountered here, including the complex problems presented by the hybrid compounds containing binational constituents. An example is German "Uraniumgewinn," "uranium yield," in which the first constituent occurs in a large number of languages in exactly the same meaning and graphic form, whereas the second constituent, "gewinn," has, I believe, not yet been loaned to other languages.

The Mechanical Abstraction of Essential Grammatical Meaning. Further research which I also carried out during the summer of 1952 revealed that it was actually very simple to devise a scheme by which a mechanical system could abstract the relevant grammatical information from the conventional written form of a source text without the necessity of human intervention. In order to achieve this, it was only necessary to arrange for a kind of filtering procedure in which identification coincided with grammatical determination. One such scheme I worked out in some detail, and it has been published in the volume *The Machine Translation of Languages.*

The Mechanical Determination of Incident Nongrammatical Meaning. Today there can be no doubt that also the syntactic analysis of source texts, or, at least, as much as is essential for mechanical translation, can be mechanized. Of the major linguistic problems facing the mechanical translation researcher this leaves only that of the mechanical determination of incident nongrammatical meaning. This problem is most intimately connected with comparative semantics, a branch of general linguistics to which I have given years of research. It is, in fact, through this study of comparative semantics that I became involved in mechanical translation.