Machine Translation of Chinese

Although perfect translation is known to be unattainable, good progress has been made in translating Chinese with a machine that has high-speed access to a linguistic "library"

by Gilbert W. King and Hsien-Wu Chang

To the Western mind the Chinese language, particularly the written language, has long been a source of wonder. Because the Westerner has known an alphabet from childhood he can scarcely believe that the strange ideographic characters of Chinese can convey the whole gamut of human thought, from philosophy to platitudes. Even without the added complication of ideographic writing, the Chinese language is so difficult that it seems most unlikely that many Americans or Europeans will ever learn it well.

Machine translation of Chinese would seem to offer the only realistic hope of giving the West ready access to the manners, achievements and aspirations of a fourth of the human race. The Indo-Chinese group of nations, with a population of about 750 million, is currently publishing in newspapers, journals and books about three billion words a year. Much less than 1 per cent of this vast output is now being translated and republished in English, French or German (undoubtedly a larger -percentage is being translated into Russian), yet nearly all of it would be of great interest to one Western group or another. Automatic translation is needed because human -translators cannot handle the volume or hope to acquire the special vocabulary needed to make good translations in a wide variety of technical fields.

Almost as soon as the electronic computer was developed some 20 years ago, its potential usefulness for language translation was recognized. Before this potential could be exploited, however, the entire technology of information processing had to be raised to its present high level. The basic problem is to find a match between natural languages and computer languages. The term "computer languages" means formal, coded instruction to a machine capable of making precisely formulated decisions. This need for stringent formalization goes far beyond the description of languages traditionally provided by linguistic scholars. As a result languages have had to be analyzed afresh by linguists willing to recognize the necessity of precise statements for a machine. In return the computer, with its capacity for subjecting an enormous amount of data to formal treatment, has served the linguist as a powerful tool for experimentation. Thus, aside from practical results, the attempt at automatic translation can hardly fail to yield new insights into the general properties of languages.

Several years ago a number of groups in the U.S. began a major effort aimed at machine translation of Russian into English. The work ranged from the highly theoretical to immediate practical efforts at computer programing. Today, under Government auspices, scientific and technical journals published in Russian are routinely translated by machine. The results, although far from perfect, have demonstrated that understandable and useful translations can be made automatically.

Work on the machine translation of Chinese was undertaken in 1960 at the University of Washington. Research programs have also been initiated at the University of California at Berkeley, Ohio State University and elsewhere. Over this same period Soviet linguists have been hard at work on machine translation of Chinese into Russian.

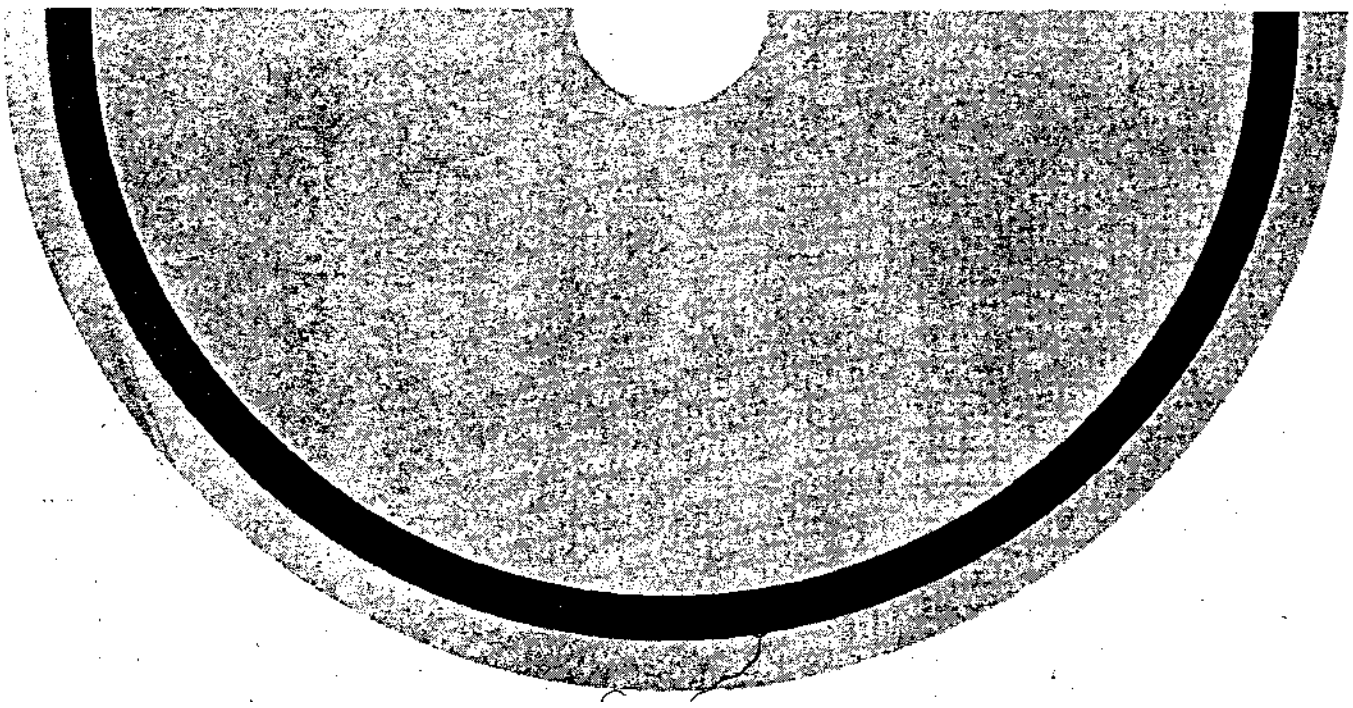
This article will describe the work of the Chinese-to-English machine translation program that the International Business Machines Corporation undertook in 1960 under a contract with the Air Force. The Chinese program has benefited greatly from ideas and machine techniques developed for an earlier Russian-to-English program. Perhaps the most important feature of that system is a photographic "memory" containing hundreds of thousands of dictionary-like entries, any one of which can be found in a twenty-thousandth of a second *[see top illustration on opposite page]*.

The need for such a large memory arose from an early recognition that a translation program could not be constructed simply by mechanizing existing grammars. It proved impossible to find any fabric of grammatical and syntactical rules that could be reduced to a manageable set of machine instructions. All spoken languages consist of a very large—in fact an infinite—set of conventions. In contrast, the operation of even the most complex machine can be expressed in terms of a relatively small set. In mechanical translation an order must be found in the larger set that makes it amenable to processing by

現在的導彈已能帶上原子彈和氫彈的彈頭，因而它是一種破壞力很大的兵器.

MODERN GUIDED-MISSILE ALREADY POSSIBLE CARRY WITH WAR HEAD OF HYDROGEN BOMB AND ATOMIC BOMB, THEREFORE IT IS ONE KIND WEAPON WITH VERY BIG POWER OF DESTRUCTION.

SAMPLE TRANSLATION of Chinese shows the present state of the machine-translation art. It was produced by methods devised by International Business Machines Corporation.

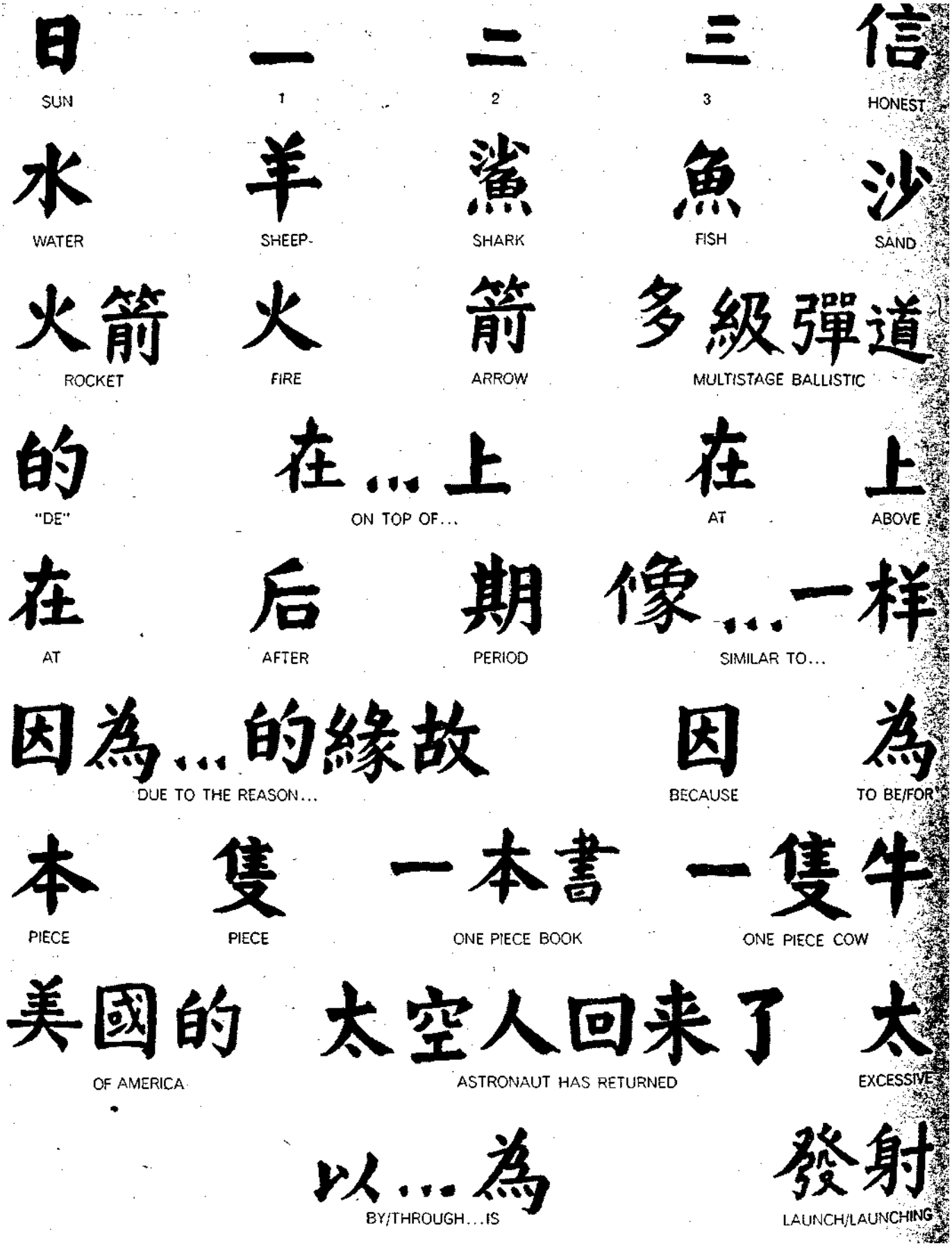


PHOTOGRAPHIC MEMORY used in Chinese-translation machine contains several hundred thousand dictionary-like entries, any one of which can be found in a twenty-thousandth of a second.

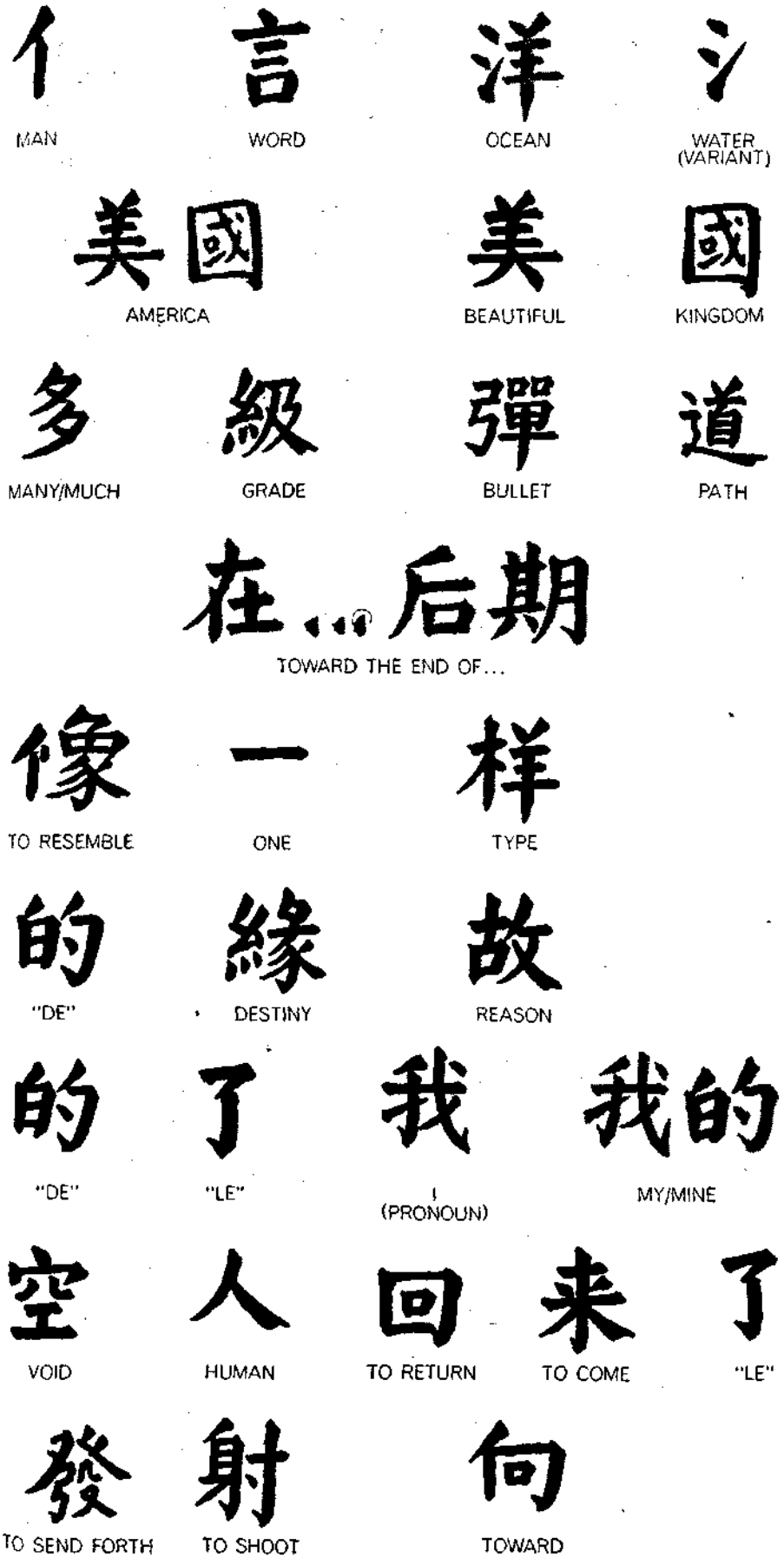
Over 50 million "bits" of information are embodied in a dot code, which forms the narrow gray bands at the edge of the disk. The disk, only half of which is shown, is printed two-thirds actual size.

```
          BAND A TRACK   8. 0416 CHINESE MASTER 2013C                PAGE      8
      8.....+]+A008+3]+A008+1]+A008+1]+A008+]
030988    O8H=366=  (T)(F38)QCC(Q4)(DP18)(Q4)(Q4)(Q4)(Q3)(Q2)(Q2)(DP9) *(841) +)
031884    O8H=3RJ=  (T)(F38)VTA(Q4)(DP18)(Q4)(Q4)2(Q2)(Q2)4(DP9) *(841) +)
031060    O8H=(Z)QJ=  (T)(F38)NLX(Q4)(DP18)(Q4)(Q4)1(Q4)6(Q4)(DP9) *(841) +)
003236    O8H=2(H=  (T)(F38)NLX(Q4)(DP18)(Q4)(Q4)(Q1)761(DP9) *(841) +)
030984    O8H=  (T)(F38)VC(Q4)(Q4)(DP18)(Q4)(Q4)(Q4)(Q3)(Q2)4(DP9) *(841) +)
002500    Q  (T)(DPO)  (M)(RX)A +)
003163    QVF=STR=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q5)(T)(Q1)17(DP9) *(841) +)
003107    QVF=A8H=KJJ=  (T)(F38)(D)(Q4)(DP18)(Q4)(Q4)(Q4)17(Q3)(DP9) *(841) +)
003194    QVF=  (T)(F38)VTA(Q4)(DP18)(Q4)(Q4)(Q1)(Q4)47(DP9) *(841) +)
031723    QRH=86K=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)17(Q1)3(DP9) *(841) +)
003082    QUJ=ZTH=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q4)(Q2)(Q1)(Q1)(DP9) *(841) +)
003081    QUJ=J8H=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q4)(Q2)(Q11(DP9) *(841) +)
003083    QUJ=ARJ=  (T)(F38)VVP(Q4)(DP18)(Q4)(Q4)(Q4)(Q2)(Q1)4(DP9) *(841) +)
003080    QUJ=2RH=1HG=  (T)(F38)QA(Q1)(Q4)(DP18)(Q4)(Q4)(Q4)(Q2)(Q3)(DP9) *(841) +)
003079    QUJ=2RH=  (T)(F38)XVN(Q4)(DP18)(Q4)(Q4)(Q4)(Q2)1(Q2)(DP9) *(841) +)
003184    QUJ=  (T)(F38)VTA(Q4)(DP18)(Q4)(Q4)1(Q3)5(Q3)(DP9) *(841) +)
030474    QUE=8RG=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q4)474(DP9) *(841) +)
030279    Q4Y=SEH=  (T)(F38)QA(Q1)(Q4)(DP18)(Q4)(Q4)(Q4)(Q1)7(Q3)(DP9) *(841) +)
030810    Q1H=C6H=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q4)(Q2)(Q4)(DP9) *(841) +)
032090    Q1H=  (T)(F38)AA(Q4)(Q4)(DP18)(Q4)(Q4)(Q1)(Q4)5(Q4)(DP9) *(841) +)
031729    Q1D=7DJ=  (T)(F38)QA(Q1)(Q4)(DP18)(Q4)(Q4)17(Q1)(Q3)(DP9) *(841) +)
031149    QKH=LKH=  (T)(F38)AA(Q4)(Q4)(DP18)(Q4)(Q4)114(Q3)(DP9) *(841) +)
031868    QKH=3YG=6MG=  (T)(F38)NND(Q4)(DP18)(Q4)(Q4)1(Q2)6(Q2)(DP9) *(841) +)
031354    QKH=3YG=  (T)(F38)NND(Q4)(DP18)(Q4)(Q4)(354(DP9) *(841) +)
031150    QKH=8GH=  (T)(F38)VVU(Q4)(DP18)(Q4)(Q4)11(Q4)(DP9) *(841) +)
032727    QKH=E8H=A9G=4HG=  (T)(F38)NN(Q4)(DP18)(Q4)(Q4)(Q1)(Q1)7(DP9) *(841) +)
031860    QKH=E8H=  (T)(F38)NND(Q4)(DP18)(Q4)(Q4)(Q2)6(Q4)(DP9) *(841) +)
030237    Q3O=22T=  (T)(F38)AA(Q4)(Q4)(DP18)(Q4)(Q4)(Q4)(Q1)37(DP9) *(841) +)
003056    Q3H=HYH=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q4)4(Q3)4(DP9) *(841) +)
003095    Q3H=IUG=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q5)(Q2)(Q4)(DP9) *(841) +)
030626    Q3H=88G=  (T)(F38)AA(Q4)(Q4)(DP18)(Q4)(Q4)(Q5)6(Q1)6(DP9) *(841) +)
031771    Q2Y=A6I=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q4)771(DP9) *(841) +)
031890    Q2H=6EI=  (T)(F38)XVN(Q4)(DP18)(Q4)(Q4)1(Q2)(Q3)(Q4)(DP9) *(841) +)
032338    Q2H=8RG=  (T)(F38)NN(Q4)(Q4)(DP18)(Q4)(Q4)(Q1)33(Q2)(DP9) *(841) +)
002502    Q  (T)(DPO)  (M)(RX)A +)
030029    PMH=ZKH=  (T)(F38)NLP(Q4)(DP18)(Q4)(Q4)(Q4)(Q4)(Q1)(Q3)(DP9) *(841) +)
030059    PMH=6LY=  (T)(F38)NNO(Q4)(DP18)(Q4)(Q4)(Q4)(Q4)5(Q3)(DP9) *(841) +)
032405    PMH=1LH=21F=  (T)(F38)NLN(Q4)(DP18)(Q4)(Q4)(Q1)4(Q4)5(DP9) *(841) +)
```

ENTRIES IN MEMORY DISK are shown as they appear when decoded and printed out in type. The number at the extreme left is an acquisition, or reference, number. This is followed by a triplet code representing one to four Chinese characters as coded by the Sinowriter, described on pages 128 and 129. The remaining symbols represent definitions and linguistic information.



GLOSSARY contains a list of Chinese characters in the order of their appearance in the accompanying text. Each Chinese character represents a monosyllable to which one or more English meanings can usually be assigned. Exceptions are certain characters, such as auxiliary words and "functives" that have no exact English counterpart. Many Chinese words are formed by combining two or



more characters. In that case the glossary usually shows the multicharacter word, or term, first (in *gray),* with its meaning, and after it the meaning of the individual characters. Frequently the individual meanings provide a clue to the meaning of the compound.

the smaller. To the extent that one cannot regularize aspects of the larger set one must have a way of mechanically storing them. This is the function of the large photographic memory.

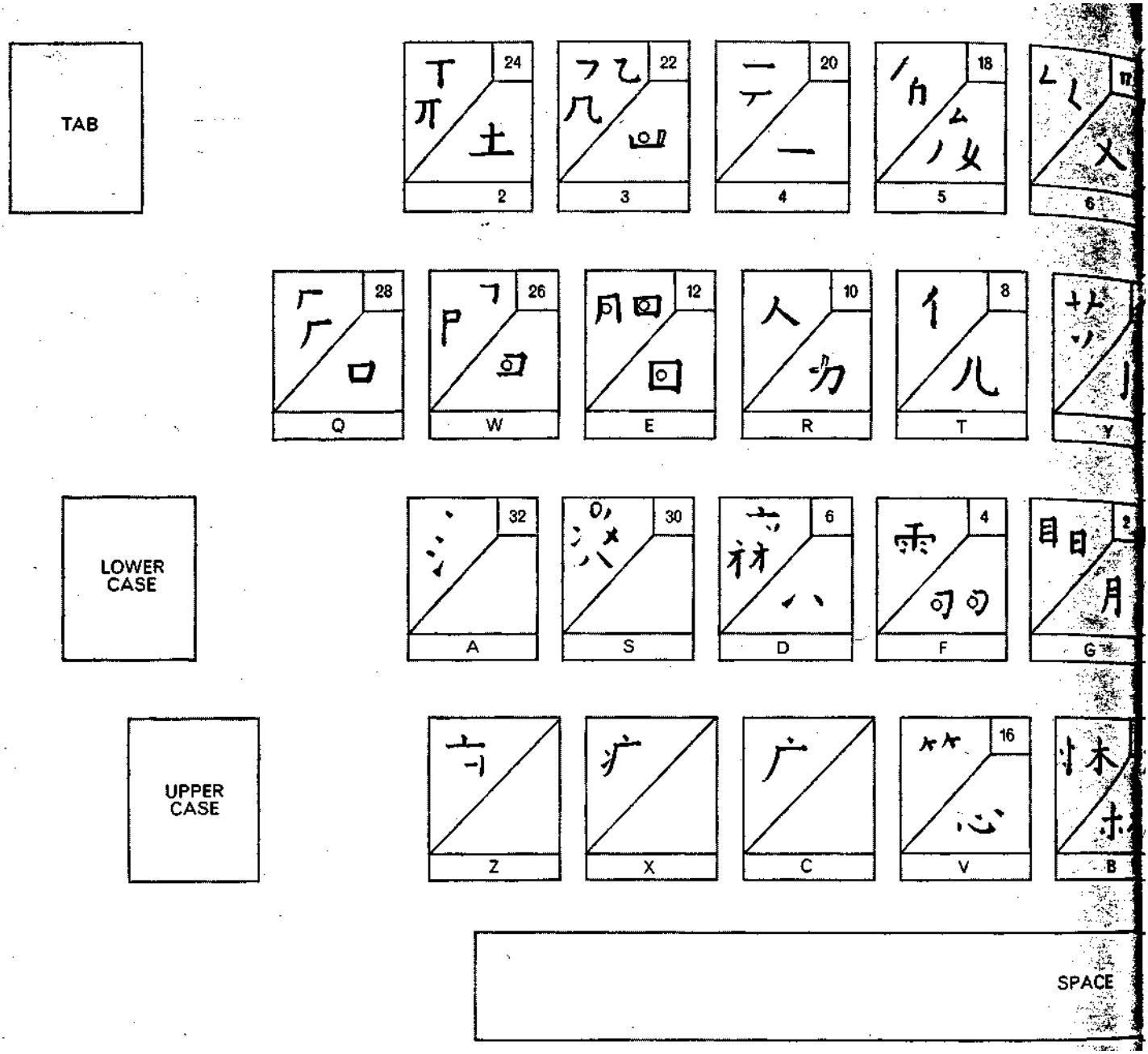
This memory contains exhaustive lists covering the many ways in which specific words can function in a sentence. It covers all the common ambiguities, transpositions and hiatuses in word order, idiomatic expressions and hundreds of special cases of the sort that make life so difficult for students learning a foreign language. In its analysis of a typical Russian .sentence the machine may refer to this combination dictionary-grammar several hundred times. The analysis is not performed on the raw input sentence but on an intermediate process sentence in which each process word contains, in an appended code, the grammatical and semantic equivalents of the corresponding input word of the sentence. In the automatic analysis grammatical relations among words are established, words are reordered and translation decisions are made. On completion of this analysis the machine has the information needed to fabricate an output sentence that represents a translation of the input.

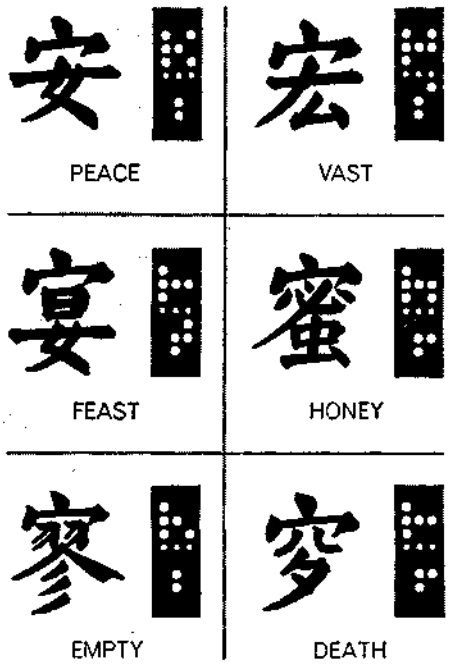
For Russian, at least, this system proved to be quite successful. Whether or not it would be of more general value in machine translation remained to be seen. Chinese, being so different from both Russian and English in its language type and family, provided an excellent test.

From the outset Chinese presents a problem not found in Russian or other alphabetic languages: the written Chinese character. The machine must be provided with a description, in code, of every character in the sentence to be translated. Moreover, to achieve speed and efficiency the code assignment must be made by a human operator at a keyboard. Machines to "read" Roman or Cyrillic characters have not yet been perfected, and a machine to read Chinese characters is nowhere in sight.

The number of characters used in the average Chinese newspaper is about 4,000. Literary and technical writing may contain 8,000 to 12,000 different characters. Chinese characters originated as stylized pictures, or hieroglyphs. As they evolved, the pictorial aspects became less pronounced, but they did not entirely disappear. They reached approximately their present form more than 2,000 years ago.

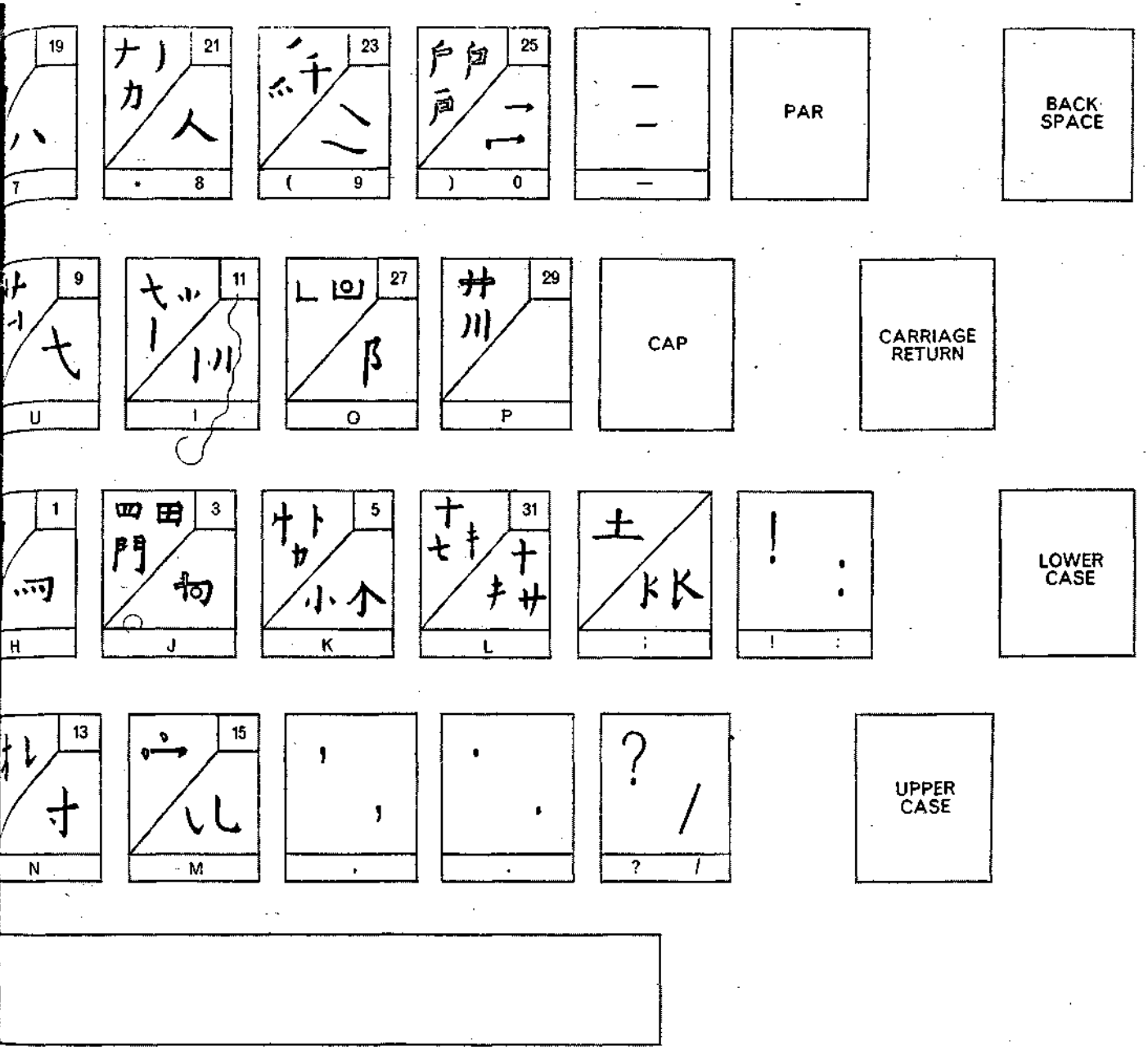
The scholar who has most recently reviewed their classification, Y. R. Chao





SINOWRITER KEYBOARD provides a means for encoding Chinese characters on standard Flexowriter tape. The designs on the keys represent characteristic features found in the top and bottom of Chinese characters. To encode a character the operator must strike three keys. The first represents a design found in the upper portion of the character. The second represents a design in the lower portion. Together these keys activate a display device that shows the operator a family of characters, all of which incorporate the same upper-half and lower-half features. Such a family of six characters is shown at bottom left on these

of the University of California at Berkeley, describes six categories. The first is made up of pictographs, which in their more ancient form were mere pictures of concrete objects. Thus 日, the symbol for "sun," is derived from an ancient form consisting of a dot in the center of a circle. In the second class are simple ideographs in which concepts are treated schematically; for example, the digits "one," "two," "three" are written 一, 二, 三. The third class contains complicated ideographs whose meaning is the combination of the meaning of their parts. For example, "honest" is written 亻言, which is derived from 亻, "man," and 言, "word." In the fourth class are loan characters, which take on the forms of other characters but have entirely different meanings. The fifth class is made up of derivative characters. The sixth class consists of phonetic compounds, which are generally formed with two component parts; one suggests the true meaning and the other indicates pronunciation. Thus 洋, "ocean," is formed from 氵, which is a variant of the character 水, "water," and 羊, which means "sheep." This indicates that 洋 has the same pronunciation as 羊 alone. (En-



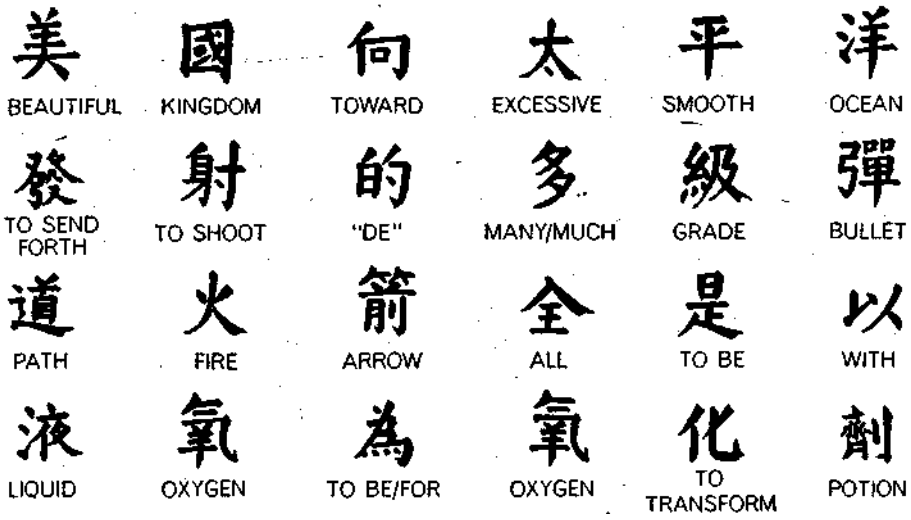
two pages. They would appear on the display if the operator struck the "M" key, for upper configuration, and the "5" key, for lower configuration. To designate the first character in the family, the character meaning "peace," the operator would strike the key with "1" in the upper right-hand corner. This is also the "H" key on the conventional Flexowriter keyboard, and the six-digit code for "H" would appear on the punched tape. Thus the complete code for peace is M5H, which is shown in punched form next to the character. The codes for the other members of the family are M5G (vast), M5J (feast), M5F (honey), M5K (empty) and M5D (death). On some keys color is used to indicate portions of a character that need not appear in the character being sought. A colored circle indicates the position of various alternative strokes. In its present form the Sinowriter contains 6,500 characters.

larged versions of these and other characters in the text can be found either in the glossary on pages 126 and 127 or in the sample sentence on the next page.)

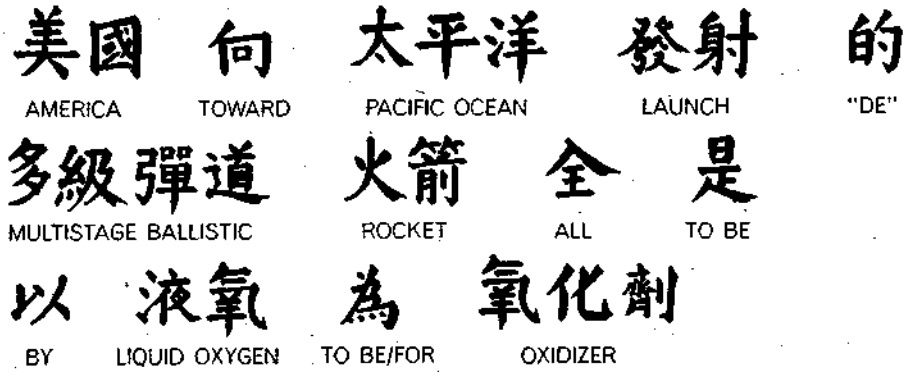
Chinese characters have also been classified according to their component parts into radicals. A character is usually made up of two elements: a radical that suggests the meaning and a phonetic that gives a clue to the pronunciation. For example, 鯊 (shark) is made up of the radical 魚 (fish) and the phonetic 沙 (sand). According to the *Kang Hsi* dictionary there are 214 radicals.

The problem of encoding Chinese characters for machine processing could be solved in any number of ways. A straightforward way to provide a code for, say, 10,000 characters would be to build a large keyboard containing 10,000 keys in a 100-by-100 array. Each key, when pressed, would punch a unique sequence of holes in paper tape. If a simple binary code ("hole" or "no hole") were used, a sequence of 14 places would be needed. (A 14-place binary code would actually be long enough to specify 16,384 characters; a 13-place code could specify only half that number.) It is obvious that even a Chinese scholar would need months to master such a keyboard and that his ultimate coding rate would never be very high.

We needed a keyboard that could be learned fairly quickly by people who are not necessarily able to read Chinese. The problem is basically one of devising a scheme for indexing characters so that they can be found readily. The problem had been much studied because it is fundamental to telegraphy, to designing a typewriter and even to selecting the presentation of characters in a Chinese dictionary. Most of the



CHINESE SENTENCE consists of strings of characters with no intervening spaces. The individual meanings of the 24 characters in this sentence are indicated. By using the "principle of the longest match" the translation machine identifies characters that should be combined to form words or word groups. The result of this processing step is illustrated below.



AFTER GROUPING OF CHARACTERS the machine finds that the sample sentence *(above)* contains 12 words or word groups ("lexical units"). The two characters "by"... "to be/for" form a discontinuous constituent that is regarded as one lexical unit for processing.

indexing schemes exploit specific details in the way Chinese characters are drawn.

We have investigated a geometric-recognition scheme that the Chinese author and scholar Lin Yutang had devised for a typewriter. A prototype machine, called the Sinowriter, was developed jointly by IBM and the Mergenthaler Linotype Company for the Air Force. In using the Sinowriter the operator is required to recognize particular shapes in the upper and lower portions of the character.

The present machine has a vocabulary of 6,500 characters, which can be expanded to 16,000. The 6,500 characters are classified into about 1,000 families according to their upper and lower configurations. The task of the operator, in observing a particular character, is to decide which of 36 upper configurations it contains and which of 30 lower configurations. The operator then presses one key to represent the upper and another to represent the lower. The two keys activate a mechanism that projects onto a screen the whole family of characters sharing these particular configurations. The family may contain only one member or as many as 16. Each member of the family is numbered from one to 16, and the operator can easily identify the one that matches the desired character in the Chinese text. He enters the appropriate number by striking a numbered key. For each character, then, the operator strikes three keys: one for upper configuration, one for lower and one for number within a family. Each stroke produces a six-hole (or six-bit) code on paper tape, so that regardless of simplicity or complexity all characters are represented by a group of three six-bit codes. Modifications in the system will undoubtedly be made on the basis of experience.

(It is perhaps obvious that the Sinowriter could be adapted to the job of setting type by machine. Evidently no typesetting machine yet exists in China, but the Japanese are known to have developed experimental models for their language, which uses characters similar to those of Chinese.)

The punched tape produced by the Sinowriter represents the raw input to the translating machine. Again Chinese presents a special problem. Except for punctuation, the printed Chinese sentence is a stream of characters without spacing. If each Chinese character represented a single word, there would be no problem, but, as we have seen, many words are composed of two characters, and combinations of three and four characters are common in literary and technical Chinese. Each character is a ,' monosyllable to which one or more English meanings can usually be assigned. And the meaning of a combination of characters is often not obvious from its elements.

For example, 美國 is the Chinese word for "America"; taken alone, 美 means "beautiful" and 國 means "kingdom." The two characters 火箭 together stand for "rocket"; individually 火 means "fire" and 箭 means "arrow." The words "multistage ballistic" are expressed in Chinese by four characters, 多級彈道, which individually represent "many," "grade," "bullet," "path."

If the Sinowriter were operated by a Chinese scholar, he could indicate the appropriate character groupings, but in his absence the groupings must be discovered by the translation machine. This is made possible by the "principle of the longest match." The dictionary of characters " stored photographically in the translation machine's memory is so ordered that any given character is presented first in groups that have a collective meaning. As the machine searches it finds shorter and shorter groups and finally single characters. The dictionary| is so arranged that the machine automatically finds the longest sequence that matches a given input.

To each entry is added a definition, similar to that in a conventional dictionary but much more formal and, in the long run, more definitive. The defini-^ tions, in fact, supply the intermediate grammatical and semantic information about the lexical unit, or language unit, that is basic to the whole translation process.

Once the longest lexical units have been identified, the machine's next step is to see how the units are related. A basic feature of modern languages is that lexical units do not necessarily make sense if they are simply translated in the

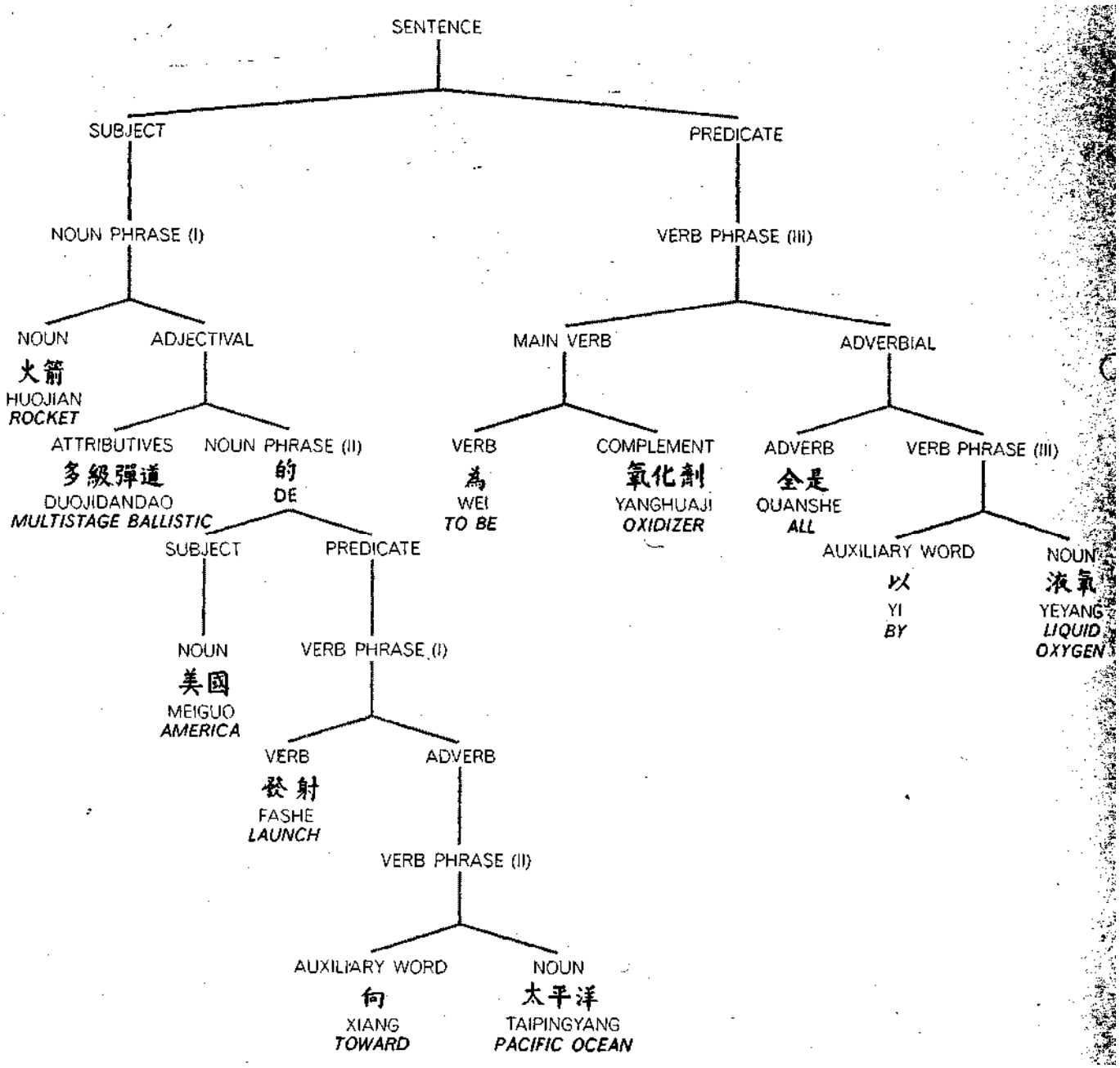
sequence in which they originally appear. The units have special relations to each other, and frequently the linkage between words or phrases is not between adjacent units but over a span of units. It turns out that all grammatical linkages can be described by a tree structure.

The machine is enabled to create a tree structure by means of "tags" it finds attached to entries in the dictionary and in the tables of grammatical rules. These tags direct the machine from one entry to another and provide, in effect, a dynamic program for processing each input sentence according to its own lexical components. The net result of this sequential look-up process is to assign a label to each lexical unit indicating its position in a tree structure. The fact that a tree has only one trunk, regardless of the number of branches, is equivalent to the recognition of a string of words as a sentence. Although any child, English-speaking or Chinese, can recognize a sentence, machines find the task extremely difficult.

The reason for this difficulty is twofold. First, all words are subtly different in function, even though the differences are often not too important. The traditional classification of English words into the eight parts of speech is a great oversimplification. The possible linkage any given word may have covers a broad spectrum and can vary almost from sentence to sentence. To provide a translation machine with a formal list of just the most common linkages is a formidable job. The second grave obstacle to sentence recognition is the familiar and rather strange fact that virtually every word can be translated in more than one way. The upshot of this twofold difficulty is that ambiguities arise in almost all cases, so that many alternative tree structures are possible on blind application of the "rules of grammar," however elaborate.

One way to handle this problem is to have the machine discover "points of entry" into a sentence. For example, the English word "the" invariably begins a noun phrase and is not linked to a preceding word. In Chinese a useful point of entry is served by the symbol $J, pronounced "de" (or, to be more precise, "duh"), which has no unique meaning of its own but combines freely with other characters to modify their meaning. For this reason linguists call the symbol a functive.

A tree structure can be built up in an unambiguous way by identifying points of entry and making linkages from these points. For this purpose the memory is supplied with permissible



TREE STRUCTURE, showing how various words are related, is a fundamental characteristic of a sentence in any language. The sentence diagramed is the one shown on page 130. Recognition of a sentence is easy for humans but difficult for a machine. In machine translation sentence recognition can be facilitated by discovering "points of entry" into a sentence. In Chinese a useful entry point is the auxiliary word pronounced "de" *(color). "De,"* which has no exact counterpart in English, *is a* functive. The pronunciation guide shown under the Chinese characters is based on the Pinyin Romanization System. No attempt has been made, however, to indicate the various tonal inflections (usually four) that can impart different meanings to the same monosyllable.

pairs of words and with a statement of the symbolic properties of these pairs, now considered as a single unit. These complex units are then used to form further pairs until the tree is established. Ambiguities can arise even then, but the procedure must be continued in the hope that a resolving clue will eventually turn up.

Sometimes these clues turn up so late that the procedure is too involved for a machine to unravel. Often clues will reside in preceding sentences in a form the machine cannot retain. Therefore one must accept the fact that real translation is impossible. The translation can nonetheless be good enough to convey as much information as the original, which itself is not always perfect.

Part of the linkage problem is the problem of the discontinuous constituent. In English familiar examples are "either... or" and "not only... but also." Such constructions are far more common in Chinese than they are in English. In many cases the constituents are nothing more than words whose component characters seldom occur sequentially in a sentence. Examples are 在 ... 上 (on top of...), 在 ... 后期 (toward the end of...), 像 ... 一样 (similar to...), 因为 ... 的緣故 (due to the reason.'..). The machine's memory has to be provided with exhaustive lists of all such examples and correct semantic translations of each.

It turns out that the discontinuous constituents, although troublesome, can serve as useful parsing devices; they usually mark phrase or clause boundaries within a sentence. The word order in such phrases or clauses, however, often requires special rearrangement. Translated by simple word-for-word substitution, a Chinese phrase might

read: "As far as atomic energy and rocketry ["de"] computation..." The proper English translation would be: "As far as the computation of atomic energy and rocketry..." A more obscure example, if it is merely translated word for word, might be: "Remove we discuss ["de"] problem in addition..." This should be translated: "In addition to the problems that we discuss ..." With proper instructions, and utilizing the discontinuous constituents as parsing boundariss, the machine can unravel such inversions and ′ produce English-sounding sentences.

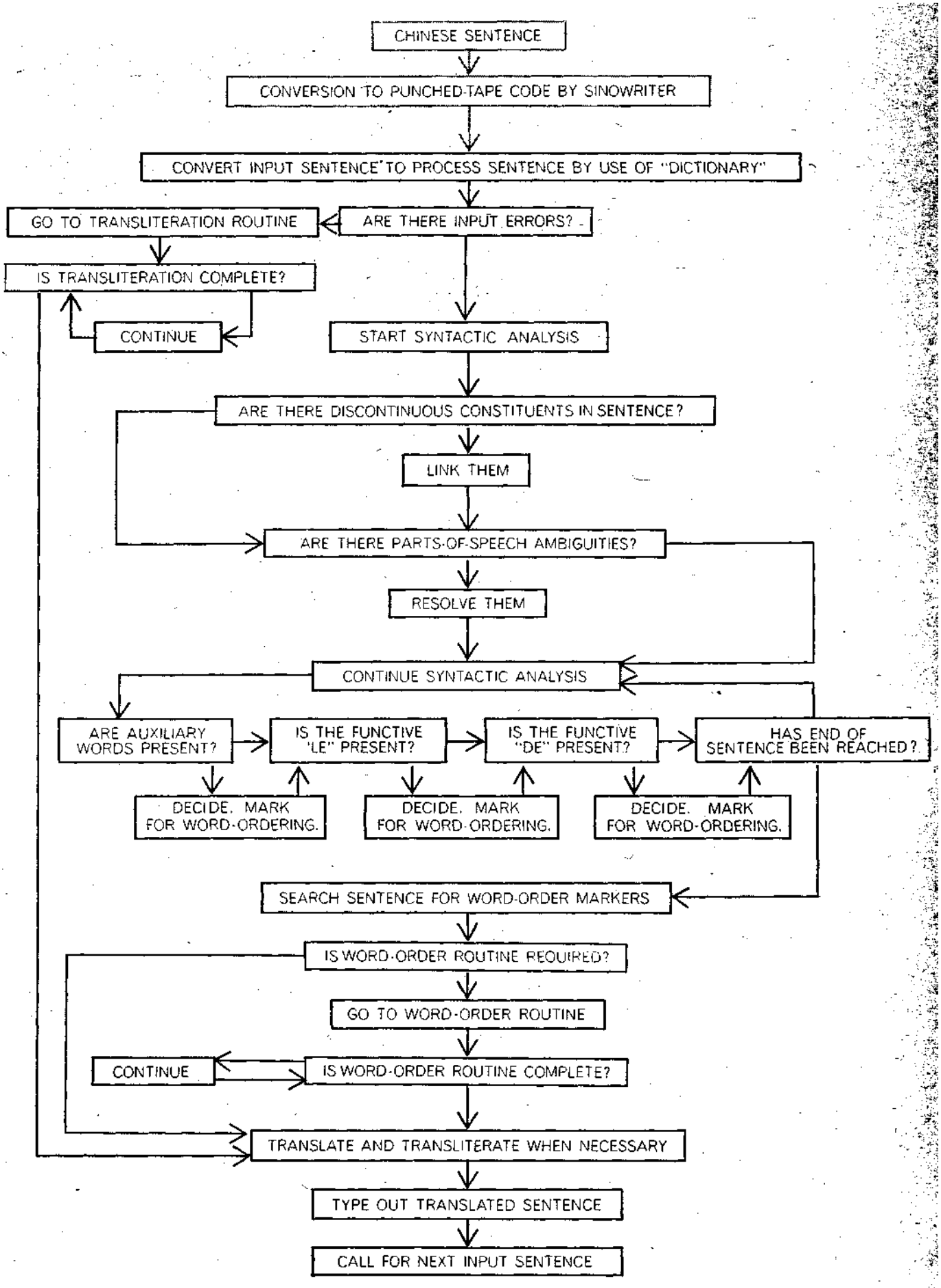
Another broad class of problems concerns the handling of auxiliary words that have no exact counterpart in English. Such words express relations among other words in sentences and serve as grammatical connections for verbs and their modifying phrases, and for noun phrases. The Chinese auxiliaries can be divided into three broad categories: verbal types, conjunctive types and nominal types. The verbal types, which are the most interesting and important, are used with verbs to express tense, voice and so on. The conjunctive types link clauses, phrases and words. Nominal types indicate such things as whether nouns are plural or singular.

Other auxiliary words are used with nouns and verbs to indicate that a measure is involved. Two such words are 本 and 隻, which can both be translated "piece," as we use the word in "one piece of paper." The first would be used in the phrase 一本書: "one piece book." The second would be used in 一隻牛: "one piece cow."

Perhaps the most versatile word in Chinese is the functive 的 ("de") mentioned earlier. Another important functive is 了 ("le"). The frequency of "de" in Chinese sentences is roughly comparable to the frequency of "the" in English, but there the similarity ends. Several examples of the use of "de" will demonstrate its ubiquity.

When "de" is used with a single noun or pronoun, it frequently imparts a possessive meaning. Thus 我 (I) plus 的 ("de") signifies "my" or "mine." Adding 的 ("de") after 美國 changes "America" to "of America." When used with a group of nouns, "de" can be translated "of," which also connotes the idea of possession. It can be used with words to form adverbs and with adjectives to form nominal modifiers. Sometimes "de" indicates the manner of action of verbs, or it may indicate that an action is completed.

The functive 了 ("le") can be used to indicate a new situation, the comple-



FLOW CHART summarizes steps required in machine translation of a Chinese sentence. The "dictionary" is the diverse linguistic data stored photographically on a disk. A colored arrow means the answer to a question is "no." A black arrow means "yes," or "proceed." The transliterative routine handles input errors and terms not in the dictionary; such items are simply transliterated.

tion of an action or even the continuation of an existing situation. A typical example of the use of 了 is in the sentence 太空人回來了: "Astronaut has returned."

Let us flow consider in some detail the steps taken by the machine in translating the .following Chinese sentence:

美國向太平洋發射的多級彈道火箭全是以液氧爲氧化劑.

Note first of all that the sentence contains 24 characters without spacing. By using the principle of the longest match the machine determines that the sentence actually contains 12 words or word groups, including a discontinuous constituent and the functive "de." (The result of this processing step is shown in the lower illustration on page 130.) A simple word-for-word translation of the sentence would read: "America toward Pacific ocean launch ["de"] multistage ballistic rocket all to be by liquid oxygen to be/for oxidizer."

The machine's next instruction is to link all discontinuous constituents in the sentence. In this case there is one 以 … 爲 (by/through/with... is/make/cause). Next the machine examines the sentence for ambiguous parts of speech and finds the word 發射, which can mean either the verb "launch" or the noun "launching." It is determined to be a verb. Following this the machine looks for auxiliary words and functives. It finds the auxiliary 向 (toward) and the functive 的 ("de"). The machine determines that "toward" heads the adverbial phrase "toward Pacific ocean." It discovers that "de" is used with the verb "launch" and, in response to a relative-clause subroutine, it inserts the word "which." Word-rearrangement markers are also tagged to all the words that are affected by this subroutine. The last step is word reordering and typing out of the translated sentence: "Multistage ballistic rocket which America launch toward Pacific ocean all use liquid oxygen as oxidizer."

This is a fair example of the results we have been able to obtain by automatic translation, using the methods briefly described in this article. Another sample translation is shown on page 124. We feel that the general usefulness of the linguistic-dictionary approach, first used with Russian, has now been demonstrated with Chinese. This is not to say that all the relevant problems are solved or will soon be solved. The goal, however, is well worth the effort, for we are seeking to remove the serious block that now exists to communications between people of the two largest language groups in the world.

LETTERS

Sirs:

The article by Gilbert W. King and Hsien-Wu Chang ["Machine Translation of Chinese," SCIENTIFIC AMERICAN, June] contains several extremely misleading statements. First, it is claimed that "scientific and technical journals published in Russian are routinely translated by machine." This simply is not true. The system that Dr. King developed while he was at IBM did indeed produce what can only euphemistically be called translations, but in a careful environment in which neither equipment nor procedure could be said to have been employed routinely.

Furthermore, the statement that "the results, although far from perfect, have demonstrated that understandable and useful translations can be made automatically" is misleading. There is no published evidence that any impartial evaluation of the output of this system has been made at all, and only a few months ago this question was a subject for inquiry by the Air Force Scientific Advisory Board, whose conclusions, one gathers, were not as sanguine as represented.

This type of misrepresentation, which has occurred quite frequently over the past decade, has unfortunately led Government agencies and others to believe that their language-translation problems are already solved. Nothing could be further from the truth, and it would be folly, for example, not to continue emphasizing the training of scientists and engineers in order to enable them to read, by themselves, foreign literature in their fields.

The statement that "it proved impossible to find any fabric of gramatical and syntactical rules that could be reduced to a manageable set of machine instructions" is not true in the world at large. Considerable progress is being made in this area at a number of research centers.

Dr. King has consistently chosen to ignore certain very serious problems. For example, the sample translation of Chinese on page 124 misleadingly suggests that machines can reliably produce a single English correspondent fora single Russian word or combination of Chinese characters. This has indeed been the case in Dr. King's Russian system, but simply because the dictionary included only one English correspondent for each Russian word, and very rarely two. This naturally has the effect of making the "translations" look good, but since most Russian words and most Chinese character combinations are hardly unambiguous, this kind of drastic oversimplification can lead to serious errors. The article similarly gives the impression that tree structures such as those displayed on page 132 can readily be obtained in an unambiguous fashion. This again is hardly the case for Russian or English, and I very much doubt that it would be the case in Chinese. There is, therefore, no guarantee that the structure developed by the machine is at all correct. Dr. King does admit in his closing paragraph that "this is not to say that all the relevant problems are solved or will soon be solved," but this mild disclaimer is hardly enough to undo the damage caused by the less responsible claims made earlier in the article.

ANTHONY G. OETTINGER

The Computation Laboratory
of Harvard University
Cambridge, Mass.

Sirs:

I would regret as much as Professor Oettinger any interpretation of the article on Chinese translation that would discourage future research and I had no intention of misrepresenting the the quality

of translation being achieved with machines at present. A real difference of opinion seems to exist on the usefulness of translations of this quality. They were found, in an operational evaluation, to be quite useful by the Government. For the past two years a contract has been in effect under which 10,000 words of Russian are translated daily. Generally these are sent in on a telephone line, processed and sent back to the Government in minutes, in what I at least regard as being a routine manner not requiring a particularly "careful" environment. Our allusion to this Government project was, perhaps, the first in open literature, **so** that Professor Oettinger's being unaware of it is understandable.

Although Professor Oettinger believes it is possible, "in the world at large," to reduce grammar and syntax to a manageable set of machine instructions, he can only support his belief with the claim that "considerable progress" is being made in this area. The fact is that progress has been slow, and a decided trend to the descriptive, or "table look-up," methods advocated in the Chinese-translation article is developing not only in machine translation but also in the field of nonnumeric processing as a whole.

I would agree with Professor Oettinger that it would be deplorable to have machine translation used as a reason for de-emphasizing the training of scientists and engineers in languages. On the other hand, I think that the ability to process languages by machine is a national requirement and that the support of research in the language-translation field by the Air Force and other Government agencies is, and will continue to be, of immense general value.

GILBERT W. KING

Vice President and
Director of Research
Itek Corporation
Lexington, Mass.