[*Mechanical Translation*, vol.3, no.1, July 1956; pp. 14-18]

An Electronic Computer Program for Translating Chinese into English

A. F. Parker-Rhodes

General Considerations

The procedure known as translation consists in the expression, through the medium of the target language, of that information which is conveyed by the text in the source language. We shall not consider here the conveyance of anything apart from "information" in the narrow sense.

We have further to consider that the information latent in the source text may not all be relevant for the purposes of the exercise. Languages differ considerably in the kinds of information which they consider as "relevant." For example, in English we cannot convey any verbal concept without at the same time adding information about when the action took place relative both to the moment of speaking and the moment of reference. In Chinese on the other hand all this extra information is regarded as irrelevant. Differences between relevant and irrelevant information are not only due to differences in linguistic habit, but may be due to the common human tendency to include irrelevant matter rather than to risk leaving out anything of importance. Theoretically, a "sufficient" translation could be defined as one which conveyed all the relevant and none of the irrelevant information. But this would be a poor aim for a computer program, (a) because when the same "irrelevancies" are present in both languages, trouble is saved by letting them pass, and (b) the rigorous pruning of, for example, English tenses, would lead to an undesirable "pidgin" effect which can in fact fairly easily be avoided.

We therefore aim instead at carrying over all the details which do not add to the operational labor involved, and as little as is necessary to inform the target text with a minimum of elegance.

Catataxis

The required information is supplied in the source text in the form of a simply-ordered series of symbols. In the case of Chinese, these symbols are "characters." I shall say nothing here as to how these characters are to be "recognized", except to emphasize that from social and moral considerations the process ought ultimately to be mechanized, and not relegated, as some have suggested, to a semi-skilled operator, which would merely replace a highly educated translator by a less developed type of worker.

The symbols in the source text, together with their ordering-relations, contain all the information available. The semantic content of these two kinds of item may be interchanged as between source and target languages. For example, we have:

Chinese	ting^{1}fang^{2}tsu	fang^{2}tsu ting^{1}
English	top house	top of house

the relation which is expressed in the Chinese text by an ordering relation, is expressed in English by the addition or omission of a word. In the case of closely-related languages such cases may be relatively few, but in general the effect of this interchangeability will be to make the distinction between "words" and "word-orderings" a nuisance. One stage of our process must therefore be to reduce all items of information, however conveyed in the source, to a common form. This stage I call "catataxy".

There are two main ways of doing this. The first is the "lexical", the second the "algorithmic". Lexical methods aim to list all the relevant forms, be they words or word-orderings, and to record for each listed item an appropriate equivalent in the target language. [An example of the application of lexical methods to catataxy is described by Mr. Richens]. On the other hand, algorithmic methods seek to prescribe rules, analogous to the rules which we learn in the elementary processes of arithmetic, whereby the significant word-orderings can be discovered and represented by numerical symbols (like those by which we convey, in the computer, the "meanings" of the separate words); and subsequently introduce further rules, to convert these symbols into others which will indicate the word order required by the target language. The method of catataxis which I have worked out is of the algorithmic type.

Metalexis

Before I describe these methods in further detail, it is necessary to consider in some detail what form those symbols will take, by which the source text is represented in the machine. These symbols will be obtained as the output of a dictionary, whose input is provided by the signs delivered to it by the reading device. Here at once we come upon what is probably the most difficult question in machine translation. How are we to sort out, from the great variety of "meanings" capable of being attached to a given word, the one appropriate to the given context? The difficulty is only partly allayed by the fact that we shall be using, in practice, restricted languages. Even in the most restricted form of Chinese, for example, chung^1 will have, among its possible meanings, "middle," "during," and "China," while fang^4 for example will require 5 or 6 "basic" equivalents.

Two considerations can be applied to choosing the appropriate meaning in such cases: contextual and grammatical. The use of contextual criteria really amounts to further restriction of our restricted language as we go along. It will consist in practice of arranging to store in the computer a series of indications of context, drawn if possible from individual words; for example, a word such as "thrilling" could be counted as excluding the context "technical papers", while a word such as "influorescence" would carry much weight in excluding, for example, "navigation". In connection with this system, each of the alternative meanings contained in a dictionary entry will carry a "key", arranged to "fit" (in a sense defined according to the elementary operating of the machine) the "lock" in which the accumulated contextual information is stored.

As regards the grammatical criterion of choice, each alternative might carry an indication of the kinds of other words it can be associated with. For example, chung^1 after a noun preceded by such verbs as tsai^4 or tao^4, and/or followed by ti(chih), may safely be rendered by "among" or (with time-words) "during". These words can themselves be identified by special signs -- "word-class indicators*. The procedure here, therefore, will involve entering at first for each word a provisional word-class indicator, indicating the W.C.I.'s of all the alternatives not excluded by the context criterion, and then, as subsequent words are read in, the provisional W.C.I.'s must be read through to see what possibilities they exclude in regard to the grammatical contexts. It may well be necessary to go through the whole sentence twice before the full range of information is brought to bear on each word.

At the end of this process, if rightly programmed, we shall have selected a single alternative for each word of the source text, and this alternative will be represented by (a) a code sign, which the output dictionary will turn into a word of the target language, and (b*)* a W.C.I, being another code sign conveying the grammatical functions possible to this word in the source language in the given context. These W.C.I.'s will provide the raw material for catataxis.

The Kind of Algorithms used in Catataxis

The program by which catataxis is carried out must begin with a master-routine which will identify the various W.C.I.'s, and direct the computer to turn to the further algorithms appropriate to each case. The identification of W.C.I.'s is done by subtraction: they are arranged in. the numerical order of their respective symbols and suitable quantities subtracted in turn from them; the computer will then recognize each by how soon the resulting number becomes negative. The processes applied to each word-class vary considerably. In each case, the objective is to build up, from the original W.C.I., a symbol which indicates not only the word-class of the word, according to an appropriate grammatical analysis of the language, but also its relations, so far as they are relevant, to the other words in this particular sentence. This symbol I have called a "taxon"; it is worthwhile to consider in some detail what form these taxa will take.

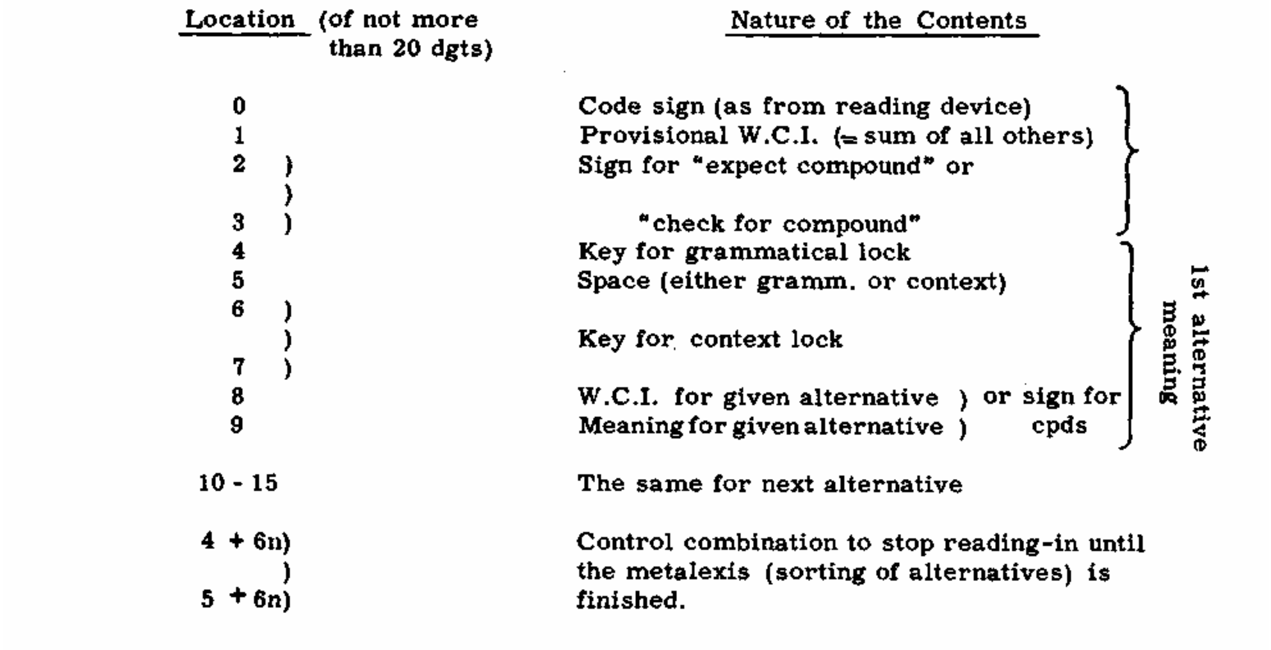
In principle, this is largely arbitrary; different methods may well *be* found convenient for different purposes. We have heard already of two possible methods of organizing sentences in mathematical terms, and the program I have proposed makes use of both "brackets" and "lattices" (or rather, chains). The only problem, in using a procedure of this type for the construction of taxa, is to select a suitable method of representing the chosen mathematical forms by the binary numerals which alone the computer can handle.

The binary representation of brackets is based in my system on the assignation of a particular binary place to each pair of brackets. Thus, in the accompanying example, in the taxa A, the square brackets[ ] enclosing the verbal group have in common, for all the enclosed words, the digits 10 in the 1st two places. The round

Table

showing the proposed arrangement of entries in the Input Dictionary

The linear order is that to be realized on the input-feed of the computer, and need not be reproduced on (say) dictionary cards.



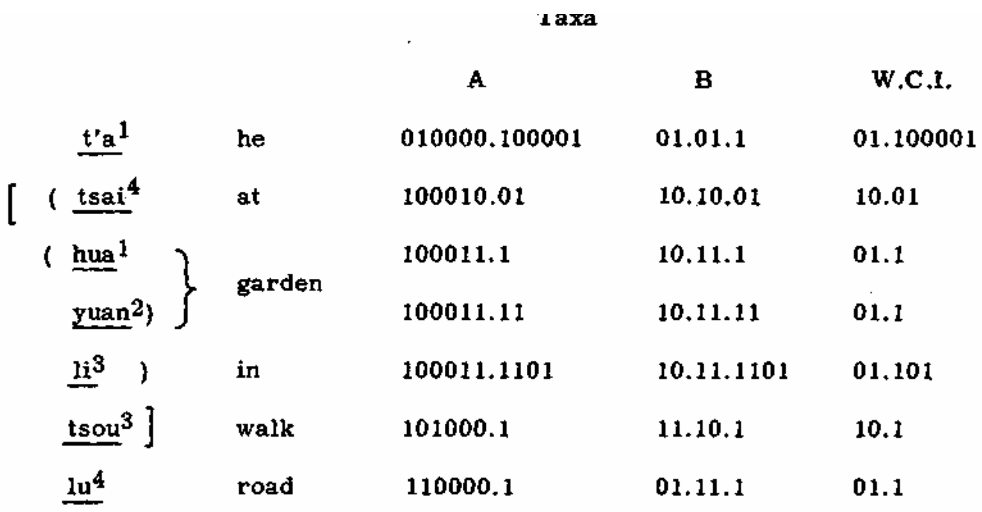
Location (of not more than 20 dgts)	Nature of the Contents	
0	Code sign (as from reading device)	
1	Provisional W.C.I. (= sum of all others)	
2)) 3)	Sign for "expect compound" or "check for compound"	
4	Key for grammatical lock	1st alternative meaning
5	Space (either gramm. or context)	
6)) 7)	Key for context lock	
8	W.C.I. for given alternative) or sign for	
9	Meaning for given alternative) cpds	
10 - 15	The same for next alternative	
4 + 6n)) 5 + 6n)	Control combination to stop reading-in until the metalexis (sorting of alternatives) is finished.	

brackets, enclosing the "complex group" (Halliday) qualifying the verb tsou^3, have in common the additional 3 digits 001; the small brackets containing the compound hua^1 yuan^2 have a further 11, which they share with their postpositive noun li^3 (in practice, such a compound as this would be separately entered in the dictionary). In this system A (which is not the one finally adopted) one can further perceive that the relation between verb and postverbal noun is indicated by the change of 01 into 11 not only at the level of the main sentence (in the 1st two binary places), but also in the subsidiary group (in the 5th and 6th places). This, in practice, is a quite unnecessary refinement; it is possible to work out the structure of all sentences completely without this information, and to abandon it makes possible much shorter taxa and simpler programming.

I therefore turned from the system exhibited in A to that of B. Here only the smaller brackets are retained, the larger brackets being replaced by a pattern of "chains". These are represented by prefixes, in which words belonging to one chain have a 1 in a prescribed position. In the example, the main-sentence chain is represented by a 1 in the second place of the prefix, and the complex-group chain by a 1 in the first place. The word tsou^3 at which the two chains join has a 1 in both places, thus showing the structure of the sentence just as clearly and much more economically than by the bracket-notation.

Having decided on the representational principles to be used in our taxa, we have to devise the necessary algorithms to derive the required binary forms from the given series of W.C.I's. This involves, first, an appropriate method of predetermining the W.C.I.'s, and, second, a set of routines for distinguishing the various groups of words which require to be recognized in the taxa. It will be noticed that in our examples the W.C.I.'s themselves form generally the last part of the finished taxon, the earlier digits being added by the algorithms. [The words yuan^2 and li^3 are exceptions, since their endings 1 and 101 receive an extra 1 to show that yuan is the second element of a compound] .

To show the sort of form our algorithms take, this last is an appropriate example.
First, when we find any taxon assuming a form identical with its predecessor, then the required algorithm is called in. Thus, at an appropriate stage, we arrange for the taxon to be subtracted from its predecessor; if the result is 0, the



			Iaxa		
			A	B	W.C.I.
	t'a^{1}	he	010000.100001	01.01.1	01.100001
[	(tsai^{4}	at	100010.01	10.10.01	10.01
	(hua^{1}	garden	100011.1	10.11.1	01.1
	yuan^{2})		100011.11	10.11.11	01.1
	li^{3})	in	100011.1101	10.11.1101	01.101
	tsou^{3}]	walk	101000.1	11.10.1	10.1
	lu^{4}	road	110000.1	01.11.1	01.1

N.B. The points are entered for ease of reading only; in the computer each digit has its fixed place and such aids are not needed.

taxon stands and is entered in the place of its W.C.I.; but if the result is 3420, we have to arrange (i) to find the last 1 in the next taxon (or the last 101 if the W.C.I. has this ending), (ii) to add a 1 in the next binary place. The taxon thus amended must be substituted for its W.C.I. In most cases, we have to add the new digits at the beginning, and to facilitate this the digits forming the W.C.I. are placed in such a position that they do not have to be shifted at all during the formation of the taxon. Often, however, a taxon has to be altered in the light of subsequent words of the sentence.

Anataxis

When all the operations required in Catataxis have been completed, all the W.C.I.'s supplied in the original input have been replaced by taxa. Each taxon is thus followed, in the storage locations of the machine, by a code sign representing its chosen "meaning" in the target language. Thus every significant feature of the given sentence, whether a word or a word-ordering, is now represented by a binary numeral. This series of signs has now to be so manipulated as to indicate correctly the order of words required in the target language.

It might in some cases be possible so to arrange the system of taxa so that they should give, by their own numerical order, the order of words ultimately required. However, this would necessitate the use of a different system of catataxis for each target language as well as for each source language, and also the algorithms required would be more complex than they need be. Thus, it is convenient to use a separate set of algorithms to alter the taxa, so as to achieve the required re-ordering.

This set of algorithms I call Anataxis, since it puts together again that which catataxis takes to pieces. (If the procedure is based on lexical methods, no separate stage is required for anataxis). As regards programming, it is simpler and shorter than Catataxis, and presents no special problems, at least as between Chinese and English which have rather similar word-orders; the main points are that in English the qualifying phrases, of the kind which in Chinese end in ti^{4} or chih^{1}, are placed after the word qualified instead of before, and that adverbs can always (though if style is to be sought, should only sometimes) follow their verbs.

In the example given above, the group in the outer round brackets needs to be placed at the end of the sentence, and this would be achieved in my program by (i) spotting it as a qualifying group (by the sequence of prefixes 01,10,11,01, separating 10,11 as the required group) and (ii) altering these prefixes so as to read, in this case, 01,11,10 (the 11 covering both the 10 and 11 of the original sequence). In other cases, other parts of the taxa must be altered; e.g.:

man^{4}	10.001		slowly	10.101
man^{4}	10.0011	becomes		1011
tsou^{3}	10.1		walking	10.0
cho^{1}	10.101			10.001

which, on arranging in numerical order, gives "walking slowly". The necessary change consists in interchanging 0 and 1 in the third place (of those here represented) from the left.

Anaptosis

When the target language is inflected (unless the inflections have fairly exact correlates in the source language) a further stage is required after Anataxis, in which the required inflections are added to the otherwise incomplete word-forms. With Chinese as the source language no assistance at all is provided in this direction, as this language is entirely uninflected. With English as the target, the difficulty is increased by the related (but logically distinct) circumstance that the required inflections mostly express logical categories which Chinese usually ignores, such as number and tense.

In my programming essays hitherto I have been content with rather crude solutions to the problems of anaptosis. Thus, I have suggested inserting "the" before all nouns where the Chinese gives no indication to the contrary (such as is afforded for example by ko^{4}, chih^{1}, etc.). Likewise, I have expected that an appropriate "blanket" tense would be acceptable in most "restricted" contexts; for example, in scientific papers, all facts may be put in the past simple, and all opinions and hypotheses in the present. The insertion of plurals can be based on the presence of particular key words. As regards case, the only distinction which appears in written English is the genitive -s, which I propose to replace everywhere by "of".

These elementary expedients would hardly serve for a more highly inflected target language, and for these anaptosis would probably have to be combined with anataxis in a single but relatively complex program.

Output

What is left in the storage of the computer when the stages of catataxy, anataxy, and anaptosis have been completed is a sequence of "words" in the order left by the anataxis routine, each of which consists of a taxon and a "meaning". The latter will have been modified so as to include sufficient information to determine the inflectional forms required, (though in a highly-inflected target language the space needed for this may be too much to be accommodated in the same location as the main "meaning" code-sign).

The taxa, however, have now served their purpose and may be cleared or overwritten, so that their places could be occupied by the additional indications required,

The last stage of the process of translation may now begin: it consists in reading-out the contents of the still relevant locations, in their present order (which is that of the target language), to a suitable output dictionary which will convert the coded "meanings" directly into alphabetic signs capable of actuating a teleprinter which will write out the target text sentence by sentence. This may be done by whatever output mechanism the given computer may be filled with. Perhaps punched teleprinter tape would be the most convenient medium.

The output dictionary need not contain any of the complications of that used for input. The latter is required to carry the necessary information for metalexis, and this process cannot be put off, since it is (in general) necessary for the determination of the W.C.I.'s which are themselves necessary for catataxis. At the output stage, however, all that is required is to decode the meaning, already determined by the code-sign which the input dictionary has supplied. Therefore, the output dictionary will work on a one-to-one basis and be correspondingly simple in design.

One of the main difficulties in mechanical translation is likely to be that of checking. In mathematical computations it is a regular and usually necessary practice to include sundry checks in the main programs. The nature of the translation process precludes this possibility. The best that can be done is to examine the output to see that it is not nonsense; this is hardly a sufficient check, but it is rather unlikely that an error in the computer would be such as to lead to "sense" other than the correct sense.