[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 5: GRAMMATICAL STUDIES

THE USE OF GRAMMARS WITHIN THE MECHANICAL TRANSLATION ROUTINE[1]

G. H. Matthews
Massachusetts Institute of Technology

I assume your acquaintance with V.H. Yngve's article, "A Framework for Syntactic Translation", which appeared in the journal Mechanical Translation, volume IV, number 3. This article conceived of translation "as a three-step process: recognition of the structure of the incoming text in terms of a structural specifier; transfer of this specifier into a structural specifier in the other language; and construction to order the output text specified". The subject matter of this talk concerns the first step (recognition routine) and the third step (construction routine).

The third step of this translation scheme consists of a grammar of the output language and a construction routine which, given the specifier for some sentence will, by reference to the grammar, produce that sentence. The form of this grammar is an adaptation of a generative type of grammar outlined by Chomsky in "Syntactic Structures". The grammar is a recursive function whose values are the grammatical sentences of the language. (In any reasonable mechanical translation scheme, these sentences would be in an allomorphic representation.) This function consists of a finite list of ordered rules in which each rule applies to the results of the preceding rules. Each rule of the grammar is either obligatory or optional, and some of the optional rules are recursive. Thus, such a finite grammar as this can generate an infinite number of sentences. Furthermore, if one specifies just which of the optional rules are to be applied and just how many times, a single specific sentence will be generated. This specification of the optional rules used can be thought of as the grammatical structure of the sentence. If the output of the second step of the translation scheme--the translation

[1] This work was supported in part by the National Science Foundation; and in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research).

proper--is a description of the grammatical structure of a sentence in terms of the optional rules to be used in its generation, it is clear that a properly programmed grammar of the language can be used as an integral part, and as a major part, of the sentence construction routine.

The sentences which are the output of such a grammar can be enumerated in the following way. The optional rules of the grammar can be assigned numerical values. Those, such as the passive rule, which can be applied just once, have the value 0 if not applied in the generation of some sentence, or 1 if applied. Those rules that are recursive, and for which there is no limit on the number of times they may be used in the generation of a sentence, may have any value, 0 if it is not used at all, 1 if it is used once, 2 if it is used twice, and so forth. There are some sets of rules, all the members of which are optional, but only one of which can be used in any one sentence, such as those which provide a modal auxiliary, that can be treated together in the assignment of a numerical value. Here the value 0 means that no modal is used, the value 1 means that the first modal in the list is used, 2 means that the second one is chosen, and so on.

We now consider a number that is formed in the following way. Each place in the number corresponds to one of the optional rules in the grammar which can have a value, or to one of the sets of optional rules that has a single value associated with it. Furthermore, the numeral that is in each place of this number is one of the values that the rule or set of rules corresponding to this place may have. There is then a one-to-one correspondence between the set of all such numbers and the sentences of the language, and since given such a number, the particular sentence it represents can be generated, the number can be regarded as the specifier of that sentence.

In addition, the sentences of a language can be given an order; namely, that order which corresponds to an order that can be given to the numbers that are specifiers of sentences. This order is the numerical order in the case of those places of the numbers that correspond to non-recursive rules or sets of rules of the grammar, and the diagonal order in the case of the recursive rules. Because of this ordering of the combinations of recursive rules, the order that is given the sentences corresponds roughly to their degree of complexity; less complex sentences normally precede more complex sentences.

Having numbers that are based on sentences in this way, it is possible to speak of significant classes of sentences of a language by number; for example, all passive sentences in English would be specified by numbers which have a 1 in the place corresponding to the passive transformation. Other more complicated--but still linguistically significant--classes are easily defined; sentences represented by numbers that have a 0 in the 1st, 8th, and 61st positions, 1 in the 6th and 104th positions, and either 24, 85, or 88 in the 6th or 14th positions. It is clear that a class of sentences that is defined by designating completely each place in the number has just one member. Hereafter, these numbers, whether they represent sentences or classes of sentences, will be referred to as specifiers.

Now I shall discuss the first step of the translation scheme, the recognition routine. The work in recognition routines that has been carried out to date by both our own group and most other groups working in the fields of mechanical translation and information retrieval has been, essentially, a search for the inverse function of the grammar of the language-- that is, a function in which the arguments are the sentences of the language, and the values are the analyses of these sentences in terms of the grammar of the language. In most cases complete analyses are considered unnecessary, or even unobtainable. The strategy is usually one of searching for enough clues in the sentence to derive as much of its analysis as possible. I now suggest that certain advantages are to be had from expressing whatever analysis can be derived from these clues with a specifier of the type described above.

The most obvious advantage derives from the fact that what is translated is a structure, not a linear sequence of words. In order to translate mechanically, there must, at some point in the process, be a representation of the structure of the sentence in a notation that makes it possible to pinpoint the sentence. The kind of specifier that I have outlined above is just such a notation. However, what one is more likely to obtain by mechanical recognition procedures is something less than a complete specification of the structure of the sentence. Thus, what is recognized is a class of sentences that includes the sentence to be recognized. The kind of specifier that I have outlined above is adapted to this situation as well. And this indicates the most important advantage to be had from these specifiers. Since it is possible to specify a class of sentences, and since it is possible to generate with a grammar all of the sentences of a specified class in an order that does not repeat the same sentence, it is possible to use a generative grammar of the language to be translated to recognize the syntactic structure of its sentences. After the preliminary recognition procedures have defined a class of sentences that includes the input sentence in terms of a specifier, a generative grammar can be used to generate all the sentences of that class, keeping track of the rules that are used in the generation of each sentence. Any such generated sentence that matches word for word the input sentence would carry with it a complete grammatical analysis of the input sentence. In this way a properly programmed grammar of the language to be translated can be used as an integral part of the sentence-recognition routine.