[*Information processing: proceedings of the International Conference on Information Processing*, Unesco, Paris 1959] 163

Symposium on machine translation

Coordinator: D.G. Hays, Rand Corporation, Santa Monica, California (USA)

1. Introduction

by *D. G. Hays.*

The chief value of a symposium when it follows (as this one does) a plenary session on the same subject, is a deeper analysis of tie principal issues already introduced. The different points of view represented in the plenary session — and those of the four speakers to be heard next — scarcely permit any quick review of MT's main points. Syntax and semantics are separated to some degree by most linguistic theories, and syntax is considered by most students of MT to have a basic place in any MT program. Thus it seems reasonable, since it is not possible to discuss the whole field of machine translation in three hours, to concentrate on the problems of syntax. Three questions may lead to a discussion of the crucial syntactic issues:

1) What is the dividing line between syntax and semantics?
2) What syntactic theory and data are needed for translation ?
3) How is syntactic analysis to be programmed for automatic machines ?

The main speakers are concerned with several languages, they are members of different schools of linguistic thought, and their work on machine translation puts syntactic analysis at different levels. Some separate sentence structure analysis from other stages of translation, and attempt to complete the analysis of each sentence before going on. Others treat as much of sentence structure as necessary for the translation of a single word, and then turn to a different segment of the sentence. Whatever agreement exists under these differences should be clarified; then we can clarify the matters on which we disagree, and the reasons for disagreement.

From this beginning, we will no doubt continue our disputes more subtly, and perhaps argue more significant points than we have sometimes done in the past.

2. Syntactic information retrieval

by *P. L. Garvin (USA).*

Syntactic information retrieval is based on the "Fulcrum" approach to syntax. The essential feature of this approach is to develop a recognition routine in terms of the differential grammatical information content of syntactic units, i.e., that component of a unit which yields the maximum amount of information is considered as its fulcrum and serves as the starting point. Sentences are treated in terms of their component clauses, and the recognition routine for each clause uses the predicate as the fulcrum for the clause, and the various nouns as fulcra for the nominal clause members. By proceeding through a hierarchy of fulcra, the recognition routine accomplishes the retrieval of the necessary syntactic information in terms of the function and boundaries of the syntactic units composing each clause of a given sentence.

In the syntactic information retrieval program for Russian into English (SIRP) which is now being tested on the 704 computer, the recognition routine is linked to a command routine which establishes the unit boundaries necessary for rearrangement into English, and accomplishes the large-scale syntactic rearrangement required for the appropriate rendering in English of Russian sentence components. At the time of writing (1 October 1958), the computer code has been written for the first of the three major phases of the verbal program (which is the equivalent of a very detailed flow chart), and the key features of the code have been successfully checked on the 704. By the time of the Conference, the entire program is expected to be computer-tested.

The verbal program, which so far covers the syntax of sentences with plural intransitive and transitive predicates is written in step-sequence style. It consists of three succeeding passes of each sentence, the first two serving to establish certain necessary unit boundaries, and the third serving to recognize major clause components, set clause boundaries, and achieve rearrangement. The recognition phase of the computer program is based on a general subroutine which serves to extract recurrent comparable elements of information; a similar general subroutine is now in preparation for the command phase of the code.

It is intended to expand the present program in several directions. Analogous rules covering additional sentence types will be formulated. Passes preceding and following the present three will allow the retrieval of certain specialized grammatical information not yet included in the present recognition routine, as well as lexical information required for syntactic commands in certain special cases.

Since SIRP consists of clearly separable recognition and command routines, its recognition routine can be used, not only to generate commands for translation as is done now, but also to retrieve and store information for other purposes, such as abstracting.

3. The NBS translation method

by *Ida Rhodes (USA).*

Let me list, in increasing order of complexity, the obstacles in the path of successful MT. These are encountered when a source sentence:

1) contains words not included in the glossary
2) omits words which must be inserted in the target sentence
3) exhibits an abnormal order of occurrences
4) contains idiomatic expressions
5) contains polysemantic words
6) is grammatically incomplete
7) is syntactically ambiguous
8) contains ambiguous symbols
9) contains localisms, grammatical errors, misprints, etc.

This classification allows us to assign to each source sentence an "Index of complexity," consisting of a 9-digit binary number, ranging from $000\ 000\ 000_2$ to $111\ 111\ 111_2$ (i.e., from zero to 2^9—1). The MT method devised at NBS permits us to handle, with relative ease, source sentences of complexity up to 2^4—1. We have only recently started deliberating on strategies for attacking the remaining 5 types of difficulty, and are hoping to obtain some enlightenment concerning this part of the task during the present conference.

Our present method consists of an iteration process, which incorporates the following features:

1) A "foresight pool," to which each occurrence may contribute statements on the nature and urgency of its affinity with future (or unexplained past) occurrences.
2) A 'hindsight pool," where note is made of any occurrence which is not predicted by the foresight pool, or of any conflict in choice, when more than one choice is permitted by the latter.
3) Pigeon-holes for storing alternate syntactic interpretations of each occurrence after a choice has been made during a given iterative cycle.
4) A "chain number," starting at one, which is only raised when the foresight pool fails to account for an occurrence.

The foresight pool receives its expectations from two sources:

1) The grammar section of the routine, which deals with general rules of grammar.
2) The glossary, whose predictions are based on the peculiar tendencies of a particular source word to affiliate with other occurrences.

At the end of a given iterative cycle, the machine would embark upon another iteration, if one or both of the following situations exists:

a) The foresight pool contains unfulfilled expectations of the highest urgency.
b) The chain number is greater than one.

Being fully aware that, in certain cases, no reasonable number of iterations will yield satisfactory results, we set a limit to the number of iterations to be performed. In such a case the printing of the unsuccessful translation is preceded by a failure signal and an indication of the types of difficulty encountered. Such a case causes us less concern than the printing of a faulty translation passed off by the machine as a satisfactory achievement.

4. "Word block model" for Russian-English syntax

by *D. R. Swanson (USA).*

A method for syntactic analysis of Russian-English translation has been developed, experimentally tested, and is at present undergoing further refinement. The syntactic rules which have been developed were initially based on the "word block model" conceived by Prof. K. E. Harper. This model can be described essentially as a phrase analysis of the sentence, and represents an elementary approach to the removal of many of the ambiguities that would otherwise be present in word-for-word translation.

We shall briefly outline the mechanized procedure which is followed. The machine operates upon an entire sentence as a unit, and syntactic analysis begins after a dictionary look-up has supplied suitable grammar codes and a group of possible English equivalents. The machine first performs a morphological and "immediate context" analysis intended to supply grammar codes for those words missing from the dictionary, and to supply part of speech codes for symbols, equations, and other uninflected words. A similar analysis of homographs is carried out in order to resolve part of speech ambiguities in so far as possible. The sentence is then marked off by the machine into "nominal word blocks" or "noun phrases." The fact that adjectives and participles usually precede the nouns they modify, and that prepositions precede their noun object, is exploited. Each word block, however, is then analyzed in its immediate environment in order to take into proper account, departures from this simple structure. Such departures, are for example, attributable to relative clauses which follow the nouns they modify, and word blocks acting as modifiers within larger word blocks.

"Word blocking" itself is a preliminary step used as a basis for rules which result in the changes already discussed. The rules themselves constitute further syntactic analysis of the sentence. It is here that we take into account the variety of dependency and governing relationships that contribute to a description of the role that each word plays within the Russian sentence. Most of the necessary modifications to word-for-word translation, such as inflection, insertion, substitution of an equivalent, and change in word order are achieved through the present set of rules. The degree to which the word block model is successful in this respect is a reflection of the extent to which "local structure" and "immediate context" play an important role in Russian syntax. The model is particularly useful in disclosing the manner in which immediate context rules fail, and consequently the nature of the remedy that must be applied.

5. Work done by the USSR Academy of Sciences in the field of machine translation

by *O. S. Koulagina (USSR).*

The USSR Academy of Sciences started its work in the field of machine translation at the end of 1954. The work, covering several languages, is now being carried on in Moscow, Leningrad, Kiev, Tbilisi, Erevan and elsewhere. It represents a first step towards the solution of the wider problem of teaching machines to process information transmitted to them in one of the world's languages. Three separate types of problem are to be distinguished in this connexion:

1) elaboration of formal systems whereby a language can be described;
2) formulation of translation algorithms for the various languages;
3) study of the problems involved in encoding and programming translation algorithms.

Translation algorithms can be placed in two categories: binary and multilingual; the application of the latter implying the existence of an intermediate language. The former relates to two languages only, and the analysis of the source language is conducted on the basis of the algorithms in accordance with the grammatical categories of the target language. Multilingual algorithms, on the other hand, are prepared for groups of languages, subject of course, to the principle of separate rules of analysis and synthesis for each of them. The results of the analysis, which serve as the starting point for the process of synthesis, are given in terms of a special language known as the "pivot-language."

Binary algorithms can be divided, in turn, into two subcategories. Some of them provide for the separate analysis of each word in the text while others relate to the analysis of word-grouping. The binary algorithms in the first subcategory would include those worked out by the Institute of Precision Engineering and Numerical Calculation for translating English, Chinese, and Japanese into Russian, as well as those worked out by the Linguistics Institute for translating Hungarian and, (in collaboration with the Mathematics Institute of the Academy of Sciences), French into Russian. This type of algorithm can be summed up as follows. After consulting the dictionary and identifying idiomatic expressions, an analysis is made of the various parts of speech in the following order: first, the cases of homography are solved, and then an analysis is made of the verbs, prepositions, nouns, pronouns, participles and adjectives.

The algorithm for translating English into Russian, as worked out by the Mathematics Institute, also provides for the use of grammatical groupings. By "grouping" is meant a group of words belonging to a well-defined class and arranged in a definite order. What the author of the algorithm has done is to determine the elementary groupings of the English language and their Russian equivalents. On analysis, the English sentence is divided into groupings which are then replaced by the corresponding Russian groupings, thus making it possible to determine the structure of the Russian sentence and obtain information as to the order and form of the Russian words. The Linguistics Institute is at present engaged in formulating an intermediate language by establishing correspondences at three different levels—lexical, morphological and syntactical—between the languages covered by its algorithm. The words of this pivot-language thus consist of word-groups corresponding to each other in the original languages. Syntactical analysis is the keystone of the translation algorithms which use a pivot-language, and makes it possible to determine the syntactical relations between the words of the source language. It is performed by means of a list of grammatical groupings, and a set of rules which enables these groupings to be identified in the text. These rules are worked out for each of the languages to which the algorithm applies.

6. Discussion

Margaret Masterman (UK): In the C.L.R.U. we also have had to be content with "Dry runs." In this way we recently half-killed ourselves translating a randomly-chosen paragraph from an Italian botanical text. This was done by means of an intermediate machine language called NUDE which was constructed in order to find out how much could be achieved with a very small number of intermediate elements. NUDE has 48 elements and 2 sentential connectives. It is part of the attempt being made by us in Cambridge to make tractable the infinities which characterise both the semantics and the syntax of natural language. Until this infinity is made finite, high-quality MT will remain a dream.

E. C. Berkeley (USA): Mr. Berkeley asked the representatives of the various MT research groups to indicate the amount of material that had already been translated by their groups. The answers were as follows:

a) A. Oettinger (Harvard University, USA): About one article per week.
b) K. Harper (RAND Corporation, USA): 500,000 words.
c) P. Garvin, (Georgetown University, USA): Very little.
d) M. Zarechnak (Georgetown University, USA): 150,000 words.
e) O. S. Koulagina (Academy of Sciences, USSR): Very little.
f) M. Masterman (Cambridge Language Research Unit, UK): Very little and that only simulation.
g) V. Yngve (Massachusetts Institute of Technology, USA): Very little.
h) S. Takahashi (Electrotechnical Laboratory, Japan) An English primer used in Japanese schools.
i) D. Swanson (Ramo-Wooldridge Corporation, USA) 60,000 words.

P. L. Garvin: If an MT program consists of a recognition routine and a command routine, the two can be considered separately. It is then possible to envisage a recognition routine, the output of which is information retrieved from the source language. A command routine for any target language can then be attached to this output. Thus, by separating the recognition and command routines, a multiple algorithm can be achieved.

Margaret Masterman: The suggestion that we should compare MT programmes by finding out how much they are algorithmic and how much they depend on finite or infinite table look-up, will not do. As a philosopher, I disapprove of infinite tables. You must have an algorithm to generate an infinite sequence of any kind and in a table consisting of semantic material you have not got this There are two infinities to be contended with in MT research: the fact that the set of possible uses of words in a language is infinite, and the fact that the set of sentence-patterns in a language is indefinitely large. Unless we can have devices—and good ones—for dealing with these two infinites, MT can be shown to be impossible.

J. Poulet (Belgique): Il parait difficile d'éviter une intervention humaine se plaçant à un certain stade en vue d'éliminer les équivalents qui ont été retenus à tort dans le processus de la traduction.

Il faudrait étudier davantage la structure de chaque langue; chacun devrait chercher d'abord à traduire, dans un langage symbolique intemédiaire et à caractère universel, la pensée exacte contenue dans sa propre langue. Ce langage symbolique peut également servir à commander des opérations à une calculatrice à partir d'un texte en langage clair, donné dans n'importe quelle langue.

J. E. Farradane (UK): The thought behind the expressions of a statement in two different languages is the true intermediate language. Language is only a tool for expressing thought, and often a poor tool. I have been working on the symbolization and coding of thought in terms of its relational structures. Would not such a "metalanguage" of thought, if realizable, be the best form of intermediate language ?

I. S. *Reed (USA):* Mr. Reed asked the members of the panel whether or not they used the coding theorems of information theory to encode words, phrases and sentences in the most economical fashion in the store.

The answer to this was unanimously negative.