

## XV. MECHANICAL TRANSLATION\*

Prof. V. H. Yngve  
Prof. A. N. Chomsky

Prof. J. R. Applegate  
Dr. G. H. Matthews

R. B. Lees  
Ursula C. Dickman

### A. PRESENT STATUS OF THE RESEARCH

The framework within which we are working is presented in schematic form in Fig. XV-1. This framework has evolved after careful consideration of a number of factors. In order to solve difficult word-order and multiple-meaning problems in translation, it seems necessary to adopt techniques that are more refined and powerful than ad hoc or empirical rules. Such techniques are being sought in sentence-for-sentence translation, which relies upon an adequate linguistic description of the sentence. These descriptions are emerging as a result of recent advances in the understanding of syntactic structure. Further information on the reasoning behind the adoption of this framework will be found in references 1, 2, and 3.

Figure XV-1 represents a hypothetical translating machine. German sentences are fed in at the left. The recognition routine, R.R., by referring to the grammar of German,  $G_1$ , analyzes the German sentence and determines its structural description,  $S_1$ , which contains all of the information that is in the input sentence. The part of the information that is implicit in the sentence (tense, voice, and so forth) is made explicit in  $S_1$ . Since a German sentence and its English translation generally do not have identical structural descriptions, we need a statement of the equivalences, E, between English and German structures, and a structural transfer routine, T.R., which consults E and transfers  $S_1$  into  $S_2$ . The construction routine, C.R., is the routine that takes the structural description of the English sentence and constructs the appropriate English sentence in conformity with the grammar of English,  $G_2$ .

The general form that the grammars G should take (4, 5) is now fairly clear: A grammar is a set of rules which allows the grammatical sequences or sentences of a language to be separated from ungrammatical sequences that are not sentences. Considerations of simplicity lead to grammars that are divided into three parts: phrase structure, transformational structure, morphology and orthography. These grammars will relate the structural descriptions of the sentences of a language to the actual sentences, and are therefore appropriate for our translation procedure.

There is no known straightforward procedure for obtaining a grammar from a corpus, or sample of language. A grammar is of the nature of a scientific theory and must be discovered. Detailed attention has been given to the development of methodology for discovering grammars. Some theoretical work has been done on the general problem of discovery procedures, and some statistical techniques have been

---

\*This work was supported in part by the National Science Foundation.

## (XV. MECHANICAL TRANSLATION)

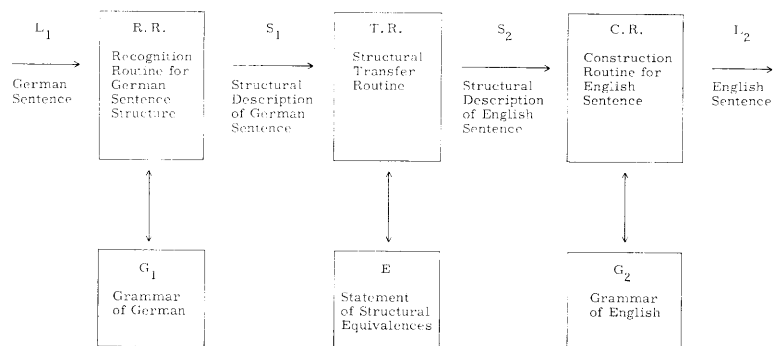


Fig. XV-1. Scheme for translating machine.

investigated which may pave the way for effective use of a digital computer for examining a very large corpus as an aid in discovering pertinent facts about a language.

A preliminary English grammar is almost finished; a German grammar is well under way. These preliminary grammars will, of course, be far from complete, but they are already complex enough to make it difficult to determine the significance, for the grammar as a whole, of changes made in some part of it. The grammars are so detailed that the use of a computer for checking them is advantageous.

Work has progressed on writing a construction routine, C.R., for the IBM 704 computer. The availability of sentence-construction routines programmed on this machine, and the ultimate availability of recognition and transfer routines, will provide an important research tool that will aid in the production of more complete and therefore more detailed grammars. One hundred thousand words of German and one hundred thousand words of English on punched tape are being made ready for research with the 704 computer.

When preliminary grammars are available, work can proceed on the important recognition routine, R.R. A straightforward approach may provide these routines. Routines using heuristic procedures, which have been investigated by those interested in theorem-proving and learning machines, also look promising. Work on grammars and recognition routines may have added significance in the field of literature search and retrieval machines.

Work on the transfer routine, T.R., awaits a preliminary statement of structural equivalences which must be based on the grammars.

V. H. Yngve

### References

1. V. H. Yngve, Syntax and the problem of multiple meaning, Chapter 14, *Machine Translation of Languages*, edited by W. N. Locke and A. D. Booth (The Technology Press, Cambridge, Mass., and John Wiley & Sons, Inc., New York, 1955).
2. V. H. Yngve, Sentence-for-sentence translation, *MT*, Vol. 2, No. 2, pp. 29-36 (Nov. 1955).
3. V. H. Yngve, The technical feasibility of translating languages by machine, *Communications and Electronics*, No. 28, pp. 792-797 (Jan. 1957); *Elec. Engng.*, 75, 994-999 (Nov. 1956).
4. N. Chomsky, Three models for the description of language, *Trans. IRE*, vol. IT-2, no. 3, pp. 113-124 (Sept. 1956).
5. N. Chomsky, *Syntactic Structures* (Mouton and Co., S.-Gravenhage, 1957).

## B. GENERATION OF NOMINAL COMPOUNDS IN ENGLISH

In any sentence-generating grammar, such as would be needed for the synthesis of target-language sentences or would form the basis of any reasonable heuristic device for analyzing source-language sentences, it will be necessary to provide for automatic generation of strongly recursive constructions that have a multiplicity of kernel-sentence sources (1). An excellent example of such constructions is the nominal compound in English, characterized by the concatenation of two or more words under a stress superfix of a primary stress (ˈ) followed by tertiary (ˋ) or secondary (ˆ); e.g., flý wheèl, tíme bòm, táx collèctor, Secûrîtíes and Exchánge Commíssíon, búrble pòint, eléctron defíciency, Mersenne nûmber, scíntillátíon còunter, fòcal pòint, and so on.

If we choose a compound of any one type, say adjective plus noun, as in hót bòx, bláckhèad, dárkròom, we note that not all members of the first category may occur together with each member of the second, but that those which do occur together are just those that also co-occur in sentences of the form The Noun is Adjective:

The man is mad                   → Madman  
The season is rainy               → Rainy season

but not:

\*The tone is hungry ..... \*Hungry tone  
\*The duchess is interstitial ..... \*Interstitial duchess.

(Note: The asterisks indicate that the forms do not occur in English.)

Since these sentences are all in the kernel, and since we wish to avoid having to state more than once the particular selections of one constituent by the other, it is reasonable to attempt to derive the compounds from the kernel sentences by means of grammatical transformations. The immediate difficulty is that not all compounds can be derived from the same type of kernel sentence, and a great variety of kernel sentences will be required for generating all types of compound.

The nominal compounds of English have been studied with a view to generating them by transformations from kernel sentences. Forty-five distinct types have been isolated, ranging in complexity from simple transformations like

The rain falls                   → Rainfall

to more complicated and perhaps more dubious transformations like

The butterfly has a tail        }                   → Swallowtail.  
The tail is like a swallow       }

Many compounds seem to be derivable by several different transformations without exhibiting any corresponding ambiguity. Thus, while the compound "ether extraction"

(XV. MECHANICAL TRANSLATION)

might be generated in two different ways – corresponding to its two different meanings:

They extract the fat with ether —————> Ether extraction (like "smallpox vaccination")

This apparatus extracts the ether —————> Ether extraction (like "arms shipment"),

the compound "ion trap" may also be generated in two different ways but seems to have only one meaning:

· This coil traps ions —————> Ion trap (like "doorstop")

This trap is for ions —————> Ion trap (like "flour sack").

Another difficulty, at present, is that in several cases a given type of source sentence appears to yield two or more kinds of compound by means of different transformations, but not all of these source sentences appear to undergo all of the transformations involved. Thus, for example, the source sentence X + V-es + N may, in some cases, be transformed into the compound V-ing + N, as in:

He drinks the water —————> Drinking water

She wears this apparel —————> Wearing apparel

and many other cases into N + V-ing, as in:

He chews tobacco —————> Tobacco chewing

They eat apples —————> Apple eating

but not all sentences underlying the second type of compound yield compounds of the first type, as in:

She reads minds —————> Mind reading but not \*Reading mind

He dodged the draft —————> Draft dodging but not \*Dodging draft.

Finally, exact specification of all of the required grammatical transformations necessarily awaits the preparation of a complete grammar of English.

R. B. Lees

References

1. For explanation of linguistic terms see Quarterly Progress Report, April 15, 1957, p. 133, and references cited therein.