[*Mechanical Translation*, vol.4, no.3, December 1957; pp. 59-65]

A Framework for Syntactic Translation †

V. H. Yngve, Massachusetts Institute of Technology, Cambridge, Massachusetts

Adequate mechanical translation can be based only on adequate structural descriptions of the languages involved and on an adequate statement of equivalences. Translation is conceived of 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 of the output text specified.

Introduction

THE CURRENT M.I.T. approach to mechanical translation is aimed at providing routines intrinsically capable of producing correct and accurate translation. We are attempting to go beyond simple word-for-word translation; beyond translation using empirical, ad hoc, or pragmatic syntactic routines. The concept of full syntactic translation has emerged: translation based on a thorough understanding of linguistic structures, their equivalences, and meanings.

The Problems

The difficulties associated with word-for-word translation were appreciated from the very beginning, at least in outline form. Warren Weaver[1] and Erwin Reifler[2] in early memoranda called attention to the problems of multiple meaning, while Oswald and Fletcher[3] began by fixing their attention on the word-order problems — particularly glaring in the case of German-to-English word-for-word translations. Over the years it has become increasingly clear that most, if not all, of the problems associated with word-for-word translation can be solved by the proper manipulation or utilization of the context. Context is to be understood here in its broadest interpretation. Contextual clues were treated in detail in an earlier article.[4] The six types of clues discussed there will be reformulated briefly here. They are:

1) The field of discourse. This was one of the earliest types of clues to be recognized. It can, by the use of specialized dictionaries, assist in the selection of the proper meaning of words that carry different meanings in different fields of discourse. The field of discourse may be determined by the operator, who places the appropriate glossary in the machine; or it may be determined by a machine routine on the basis of the occurrences of certain text words that are diagnostic of the field.

---

† This work was supported 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); and in part by the National Science Foundation.

1. Warren Weaver, "Translation," Machine Translation of Languages, edited by Locke and Booth (New York and London, 1955)

2. Erwin Reifler, "Studies in Mechanical Translation No. 1, MT, " mimeographed (Jan. 1950)

3. Oswald and Fletcher, "Proposals for the Mechanical Resolution of German Syntax Patterns, " Modern Language Forum, vol. XXXVI, no. 2-4 (1951)

4. V. H. Yngve, "Terminology in the Light of Research on Mechanical Translation, " Babel, vol. 2, no. 3 (Oct. 1956)

2) Recognition of coherent word groups, such as idioms and compound nouns. This clue can provide a basis for translating such word groups correctly even when their meaning does not follow simply from the meanings of the separate words.

3) The syntactic function of each word. If the translating program can determine syntactic function, clues will be available for solving word-order problems as well as a large number of difficult multiple-meaning problems. Clues of this type will help, for example, in determining whether der in German should be translated as an article or as a relative or demonstrative pronoun, and whether it is nominative, genitive, or dative. They will also assist in handling the very difficult problems of translating prepositions correctly.

4) The selectional relations between words in open classes, i.e., nouns, verbs, adjectives, and adverbs. These relations can be utilized by assigning the words to various meaning categories in such a way that when two or more of these words occur in certain syntactic relationships in the text, the correct meanings can be selected.

5) Antecedents. The ability of the translating program to determine antecedents will not only make possible the correct translation of pronouns, but will also materially assist in the translation of nouns and other words that refer to things previously mentioned.

6) All other contextual clues, especially those concerned with an exact knowledge of the subject under discussion. These will undoubtedly remain the last to be mechanized.

Finding out how to use these clues to provide correct and accurate translations by machine presents perhaps the most formidable task that language scholars have ever faced.

Two Approaches

Attempts to learn how to utilize the above-mentioned clues have followed two separate approaches. One will be called the "95 per cent approach" because it attempts to find a number of relatively simple rules of thumb, each of which will translate a word or class of words correctly about 95 per cent of the time, even though these rules are not based on a complete understanding of the problem. This approach is used by those who are seeking a short-cut to useful, if not completely adequate, translations. The other approach concentrates on trying to obtain a complete understanding of each portion of the problem so that completely adequate routines can be developed.

At any stage in the development of mechanical translation there will be some things that are perfectly understood and can therefore serve as the basis for perfect translation. In the area of verb, noun, and adjective inflection, it is possible to do a "100 per cent job" because all the paradigms are available and all of the exceptions are known and have been listed. In this area one need not be satisfied with anything less than a perfect job.

At the same time there will be some things about language and translation that are not understood. It is in this area that the difference between the two approaches shows up. The question of when to translate the various German, French, or Russian verb categories into the different sets of English verb categories is imperfectly understood. Those who adopt the 95 per cent approach will seek simple partial solutions that are right a substantial portion of the time. They gain the opportunity of showing early test results on a computer. Those who adopt the 100 per cent approach realize that in the end satisfactory mechanical translation can follow only from the systematic enlarging of the area in which we have essentially perfect understanding.

The M.I. T. group has traditionally concentrated on moving segments of the problem out of the area where only the 95 per cent approach is possible into the area where a 100 per cent approach can be used. Looking at mechanical translation in this light poses the greater intellectual challenge, and we believe that it is here that the most significant advances can be made.

Syntactic Translation

Examination of the six types of clues mentioned above reveals that they are predominantly concerned with the relationships of one word to another in patterns. The third type — the ability of the program to determine the syntactic function of each word — is basic to the others. It is basic to the first: If the machine is to determine correctly the field of discourse at every point in the text, even when the field changes within one sentence, it must use the relationship of the words in syntactic patterns as the key for finding which words refer to which field. It is basic to the second because idioms, noun compounds, and so on, are merely special patterns of words that stand out from

more regular patterns. It is basic to the fourth because here we are dealing with selectional relationships between words that are syntactically related. It is basic to the fifth because the relationship of a word to its antecedent is essentially a syntactic relationship. It is probably even basic to the last, the category of all other contextual clues.

Any approach to mechanical translation that attempts to go beyond mere word-for-word translation can with some justification be called a syntactic approach. The word "syntactic" can be used, however, to cover a number of different approaches. Following an early suggestion by Warren Weaver,[1] some of these take into consideration only the two or three immediately preceding and following words. Some of them, following a suggestion by Bar-Hillel,[5] do consider larger context, but by a complicated scanning forth and back in the sentence, looking for particular words or particular diacritics that have been attached to words in the first dictionary look-up. To the extent that these approaches operate without an accurate knowledge and use of the syntactic patterns of the languages, they are following the 95 per cent approach.

Oswald and Fletcher[3] saw clearly that a solution to the word-order problems in German-to-English translation required the identification of syntactic units in the sentence, such as nominal blocks and verbal blocks. Recently, Brandwood[6] has extended and elaborated the rules of Oswald and Fletcher. Reifler,[7] too, has placed emphasis on form classes and the relationship of words one with the other. These last three attempts seem to come closer to the 100 per cent way of looking at things.

Bar-Hillel,[8] at M.I.T., introduced a 100 per cent approach years ago when he attempted to adapt to mechanical translation certain ideas of the Polish logician Ajdukiewicz. The algebraic notation adopted for syntactic categories, however, was not elaborate enough to express the relations of natural languages.

Later, the author [9, 10] proposed a syntactic method for solving multiple-meaning and word-order problems. This routine analyzed and translated the input sentences in terms of successively included clauses, phrases, and so forth.

More recently, Moloshnaya [11] has done some excellent work on English syntax, and Zarechnak[12] and Pyne[13] have been exploring with Russian a suggestion by Harris[14] that the text be broken down by transformations into kernel sentences which would be separately translated and then transformed back into full sentences. Lehmann,[15] too, has recently emphasized that translation of the German noun phrase into English will require a full descriptive analysis.

---

5. Y. Bar-Hillel, "The Present State of Research on Mechanical Translation, " American Documentation, 2:229-237 (1951)

6. A. D. Booth, L. Brandwood, J. P. Cleave, Mechanical Resolution of Linguistic Problems, Academic Press (New York, 1958)

7. Erwin Reifler, "The Mechanical Determination of Meaning, " Machine Translation of Languages, edited by Locke and Booth (New York and London, 1955)

8. Y. Bar-Hillel, "A Quasi-Arithmetical Notation for Syntactic Description, " Language, vol. 29, no. 1 (1953)

9. V. H. Yngve, "Syntax and the Problem of Multiple Meaning," Machine Translation of Languages, edited by Locke and Booth (New York and London, 1955)

10. V. H. Yngve, "The Technical Feasibility of Translating Languages by Machine," Electrical Engineering, vol. 75, no. 11 (1956)

11. T. N. Moloshnaya, "Certain Questions of Syntax in Connection with Machine Translation from English to Russian," Voprosy Yazykoznaniya. no. 4 (1957)

12. M. M. Zarechnak, "Types of Russian Sentences," Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies, Georgetown University (1957)

13. J. A. Pyne, "Some Ideas on Inter-structural Syntax," Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies, Georgetown University (1957)

14. Z. S. Harris, "Transfer Grammar," International Journal of American Linguistics, vol. XX, no. 4 (Oct. 1954)

15. W. P. Lehmann, "Structure of Noun Phrases in German," Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies, Georgetown University (1957)

In much of the work there has been an explicit or implicit restriction to syntactic relationships that are contained entirely within a clause or sentence, although it is usually recognized that structural features, to a significant extent, cross sentence boundaries. In what follows, we will speak of the sentence without implying this restriction.

The Framework

The framework within which we are working is presented in schematic form in Fig. 1. This framework has evolved after careful consideration of a number of factors. Foremost among these is the necessity of breaking down a problem as complex as that of mechanical translation into a number of problems each of which is small enough to be handled by one person.

Figure 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, analizes the German sentence and determines its structural description or specifier, 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 structure transfer routine, T.R., which consults E and transfers S_1 into S_2, the structural description, or specifier, of the English sentence. The construction routine, C.R., is the routine that takes S_2 and constructs the appropriate English sentence in conformity with the grammar of English, G_2.

This framework is similar to the one previously published[16] except that now we have added the center boxes and have a much better understanding of what was called the "message" or transition language — here, the specifiers. Andreyev[17] has also recently pointed out that translation is essentially a three-step process and that current published proposals have combined the first two steps into one. One might add that some of the published proposals even try to combine all three steps into one. The question of whether there are more than three steps will be taken up later.

A few simple considerations will make clear why it is necessary to describe the structure of each language separately. First, consider the regularities and irregularities of declensions and conjugations. These are, of course, entirely relative to one language.

Context, too, is by nature contained entirely within the framework of one language. In considering the translation of a certain German verb form into English, it is necessary to understand the German verb form as part of a complex of features of German structure including possibly other verb forms within the clause, certain adverbs, the structure of neighboring clauses, and the like. In translating into English, the appropriate complex of features relative to English structure must be provided so that each verb form is understood correctly as a part of that English complex.

The form of an English pronoun depends on its English antecedent, while the form of a German pronoun depends on its German antecedent — not always the same word because of the multiple-meaning situation. As important as it is to locate the antecedent of the input pronoun in the input text, it is equally important to embed the output pronoun in a proper context in the output language so that its antecedent is clear to the reader.

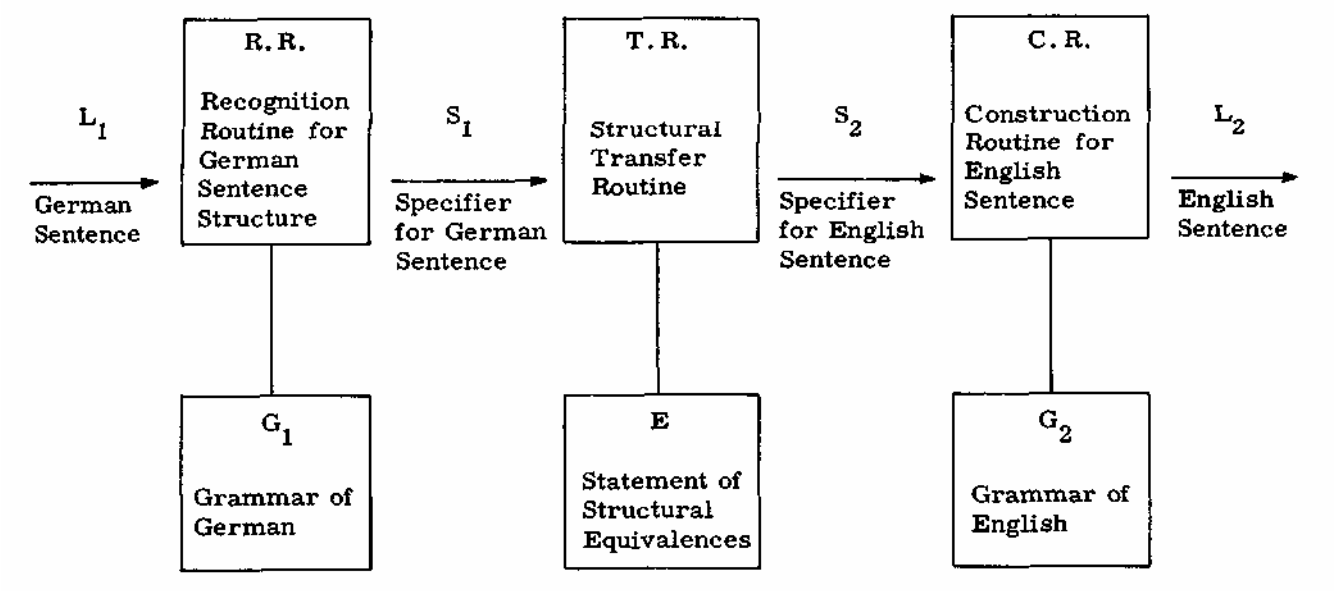
In all of these examples it is necessary to understand the complete system in order to program a machine to recognize the complex of features and to translate as well as a human translator. If one is not able to fathom the complete system, one has to fall back on hit-or-miss alternative methods — the 95 per cent approach. In order to achieve the advantages of full syntactic translation, we will have to do much more very careful and detailed linguistic investigation.

Stored Knowledge

The diagram (Fig. 1) makes a distinction between the stored knowledge (the lower boxes) and the routines (the upper boxes). This distinction represents a point of view which may be academic: In an actual translating program the routine boxes and the stored knowledge boxes might be indistinguishable. For our purpose, however, the lower boxes represent our

16. V. H. Yngve, "Sentence-for-sentence Translation," MT, vol. 2, no. 2 (1955)

17. N. D. Andreyev, "Machine Translation and the Problem of an Intermediary Language, Voprosy Yazykoznaniya, no. 5 (1957)



A Framework for Mechanical Translation

Figure 1

knowledge of the language and are intended not to include any details of the programming or, more particularly, any details of how the information about the languages is used by the machine. In other words, these boxes represent in an abstract fashion our understanding of the structures of the languages and of the translation equivalences. In an actual translating machine, the contents of these boxes will have to be expressed in some appropriate manner, and this might very well take the form of a program written in a pseudo code, programmable on a general-purpose computer. Earlier estimates[9] that the amount of storage necessary for syntactic information may be of the same order of magnitude as the amount of storage required for a dictionary have not been revised.

Construction

The Construction Routine, C.R. in Figure 1, constructs to order an English sentence on the prescription of the specifier, S_2. It does this by consulting its pharmacopoeia, the grammar of English, G_2, which tells it how to mix the ingredients to obtain a correct and grammatical English sentence, the one prescribed.

The construction routine is a computer program that operates as a code conversion device, converting the code for the sentence, the specifier, into the English spelling of the sentence . The grammar may be looked upon in this light as a code book, or, more properly, as an algorithm for code conversion. Alternately the construction routine can be regarded as a function generator. The independent variable is the specifier, and the calculated function is the output sentence. Under these circumstances, the grammar, G_2, represents our knowledge of how to calculate the function.

The sentence construction routine resembles to some extent the very suggestive sentence generation concept of Chomsky,[18] but there is an important difference. Where sentence generation is concerned with a compact representation of the sentences of a language, sentence construction is concerned with constructing, to order, specified sentences one at a time. This difference in purpose necessitates far-reaching differences in the form of the grammars.

---

18. Noam Chomsky, Syntactic Structures, Mouton and Co., 'S-Gravenhage (1957)

Specifiers

For an input to the sentence construction routine, we postulated an encoding of the information in the form of what we called a specifier. The specifier of a sentence represents that sentence as a series of choices within the limited range of choices prescribed by the grammar of the language. These choices are in the nature of values for the natural coordinates of the sentence in that language. For example: to specify an English sentence, one may have to specify for the finite verb 1st, 2nd, or 3rd person, singular or plural, present or past, whether the sentence is negative or affirmative, whether the subject is modified by a relative clause, and which one, etc. The specifier also specifies the class to which the verb belongs, and ultimately, which verb of that class is to be used, and so on, through all of the details that are necessary to direct the construction routine to construct the particular sentence that satisfies the specifications laid down by the author of the original input sentence.

The natural coordinates of a language are not given to us a priori, they have to be discovered by linguistic research.

Ambiguity within a language can be looked at as unspecified coordinates. A writer generally can be as unambiguous as he pleases — or as ambiguous. He can be less ambiguous merely by expanding on his thoughts, thus specifying the values of more coordinates. But there is a natural limit to how ambiguous he can be without circumlocutions. Ambiguity is a property of the particular language he is using in the sense that in each language certain types of ambiguity are not allowed in certain situations. In Chinese, one can be ambiguous about the tense of verbs, but in English this is not allowed: one must regularly specify present or past for verbs. On the other hand, one is usually ambiguous about the tense of adjectives in English, but in Japanese this is not allowed.

It may be worth while to distinguish between structural coordinates in the narrow sense and structural coordinates in a broader, perhaps extra linguistic sense, that is, coordinates which might be called logical or meaning coordinates. As examples, one can cite certain English verb categories: In a narrow sense, the auxiliary verb 'can' has two forms, present and past. This verb, however, cannot be made future or perfect as most other verbs can. One does not say 'He has can come,' but says, instead, 'He has been able to come,' which is structurally very different. It is a form of the verb 'to be' followed by an adjective which takes the infinitive with 'to.' Again the auxiliary 'must' has no past tense and again one uses a circumlocution — 'had to.' If we want to indicate the connection in meaning (paralleling a similarity in distribution) between 'can' and 'is able to' and between 'must' and 'has to,' we have to use coordinates that are not structural in the narrow sense. As another example, there is the use of the present tense in English for past time (in narratives), for future time ('He is coming soon'), and with other meanings. Other examples, some bordering on stylistics, can also be cited to help establish the existence of at least two kinds of sentence coordinates in a language, necessitating at least two types of specifiers.

A translation routine that takes into consideration two types of specifiers for each language would constitute a five-step translation procedure. The incoming sentence would be analyzed in terms of a narrow structural specifier. This specifier would be converted into a more convenient and perhaps more meaningful broad specifier, which would then be converted into a broad specifier in the other language, then would follow the steps of conversion to a narrow specifier and to an output sentence.

Recognition

One needs to know what there is to be recognized before one can recognize it. Many people, including the author, have worked on recognition routines. Unfortunately, none of the work has been done with the necessary full and explicit knowledge of the linguistic structures and of the natural coordinates.

The question of how we understand a sentence is a valid one for linguists, and it may have an answer different from the answer to the question of how we produce a sentence. But it appears that the description of a language is more easily couched in terms of synthesis of sentences than in terms of analysis of sentences. The reason is clear. A description in terms of synthesis is straightforward and unambiguous. It is a one-to-one mapping of specifiers into sentences. But a description in terms of analysis runs into all of the ambiguities of language that are caused by the chance overlapping of different patterns: a given sentence may be understandable in terms of two or more different specifiers. Descriptions in terms of analysis will probably not be available until after we

have the more easily obtained descriptions in terms of synthesis.

The details of the recognition routine will depend on the details of the structural description of the input language. Once this is available, the recognition routine itself should be quite straightforward. The method suggested earlier by the author[9] required that words be classified into word classes, phrases into phrase classes, and so on, on the basis of an adequate descriptive analysis. It operated by looking up word-class sequences, phrase-class sequences, etc., in a dictionary of allowed sequences.

Transfer of Structure

Different languages have different sets of natural coordinates. Thus the center boxes (Fig. 1) are needed to convert the specifiers for the sentences of the input language into the specifiers for the equivalent sentences in the output language. The real compromises in translation reside in these center boxes. It is here that the difficult and perhaps often impossible matching of sentences in different languages is undertaken. But the problems associated with the center box are not peculiar to mechanical translation. Human translators also face the very same problems when they attempt to translate. The only difference is that at present the human translators are able to cope satisfactorily with the problem.

We have presented a framework within which work can proceed that will eventually culminate in mechanical routines for full syntactic translation. There are many aspects of the problem that are not yet understood and many details remain to be worked out. We need detailed information concerning the natural coordinates of the languages. In order to transfer German specifiers into English specifiers, we must know something about these specifiers. Some very interesting comparative linguistic problems will undoubtedly turn up in this area.

The author wishes to express his indebtedness to his colleagues G. H. Matthews, Joseph Applegate, and Noam Chomsky, for some of the ideas expressed in this paper.