XI. MECHANICAL TRANSLATION*

V. H. Yngve
Irene Bellert
J. L. Bennett
Carol Bosche
J. S. Bross
Elinor Charney
Raymonde Dallaire
J. L. Darlington
D. A. Dinneen
Joan B. Galbraith
Ursula Hahn
Muriel Kannel
E. S. Klima
K. C. Knowlton
D. Lieberman
G. H. Matthews
W. K. Percival
A. C. Satterthwait
J. J. Viertel

A. SYNTACTIC PHENOMENA

Considerable research has been carried out into certain phenomena of English. It has gradually become evident that a large number of the complications of English grammar and syntax can now be understood as serving a definite function in the language. This function is to limit the language in such a way that the maximum temporary memory required for processing the language need never exceed about seven items, and to insure, in spite of this drastic limitation, that the expressive power of the language is not unduly curtailed. It seems likely that all spoken languages are subject to this limitation, that the greatly divergent complexities of different unrelated languages can now be understood, and that additional insights into language change will be forthcoming.

The particular model of sentence production adopted seemed likely to be adequate for all of English syntax. Since the model was programmable on a computer, research was initiated to prepare a computer program for the production of grammatical English sentences. It was hoped that if there were any inadequacies in the model, they would show themselves as work on the preparation of the required programmable English grammar proceeded. The initial attempt has been completed. No inadequacies have appeared.

V. H. Yngve

B. RECOGNITION ROUTINE

Work has been progressing on a general computer program that will recognize the grammatical structure of sentences of natural languages. This program scans the input sentence for those elements of its grammatical structure that are easily observed; by reference to a generative grammar of the input language, it then constructs all of the sentences of that language that have at least these structural elements, keeping track of what other elements are put into each sentence. Thus it finds out which of these constructed sentences match the input sentence word-for-word. In principle, this method of sentence recognition cannot fail to work, even though the only initially recognized structural element of the input sentence is the number of words that it has.

*This work is supported in part by National Science Foundation.

That such a program as this is both possible and practical was realized only after certain theoretical discoveries were made concerning the logical structure of the grammars of natural languages. A well-ordering, without accumulation points, which roughly corresponds to the order of "from shorter to longer" and "from less complex to more complex" can be described for the sentences of a language. A one-to-one correspondence between sentences and integers can be established, and, with a judicious choice of radix for representing these integers, sentences can be related to one another, and their grammatical structure systematically altered, by performing arithmetical operations on their corresponding integers. Results of these theoretical investigations, as well as others, are represented directly in the computer program.

The advantages of such a recognition routine are three: (a) the procedure for recognizing sentence structure is in no way dependent upon any particular language; (b) the structure of the grammar of the input language, which the recognition routine uses as a subroutine, is formally the same as the structure of the grammar of the output language, which is needed in the sentence construction program – the last step of the translation scheme that is being developed at M. I. T.; (c) grammars, with the same formal properties, which are being written by linguists in other fields than mechanical translation, can be used directly in our recognition program.

G. H. Matthews

C. COMPARATIVE SYNTAX

Work is continuing on English and Russian syntax. Within English grammar, the work begun on Negation has been expanded and includes research on more general grammatical categories. The problem that has been attacked is essentially: To what extent are more subtle similarities between existing grammatical categories symptomatic of an even higher level grammatically significant category that includes the other categories? Specifically, after investigating and describing independently negative constructions, interrogative constructions, and certain types of complementary structures subordinated in a particular way to their heads, it was discovered that there exist in all three certain common features (such as the occurrence of the indefinites EVER, AT ALL, ANY, ANYBODY, etc.) that are most easily described by postulating the grammatical similarity of these three types of construction. The particular nature of this similarity is not yet completely clear, but at the present stage of the investigation it seems very likely that the similarity in observed behavior is attributable to a common relationship between the motivating constituent (interrogative morpheme, negative operator, or phrasal head of a particular type) and the domain of the other – be it the rest of the sentence or just a complementary clause or phrase. The relationship that is common to all of these constructions, which suggests itself more and more definitively as the work progresses,

is that of a special variety of subordination.

Not only through further refinement in the description of such fine points in English syntax but also through comparison with other languages – for what we aim for is not what is peculiarly English but rather what is of general linguistic relevance – we are led to the more inclusive pursuit: To consider the type of grammatical statements that would be found in an adequate transfer grammar. In "adequate" are included not only correctness but also elegance and the quality of being revealing. The approach to the problem is expressed by the question: Given some sentence in Language A and a known adequate correspondence to it in Language B, how can this correspondence be described, and in what sense does it more nearly correspond than some other sentence in Language B? Included in the problem, of course, is the more restricted question: In what sense is word "a" of Language A the correct rendering in a given context of word "b" of Language B? I was led to these considerations of transfer grammar by the observation that with negation, interrogation, and particularly in the case of certain subordinating head words, different languages group their vocabulary in more or less the same categories – although the various more or less evident marks of the differentiation in categories vary greatly from language to language (for example, a type of quantifier is used in English where a subjunctive of the verb is found in French). It was discovered that two items of vocabulary that were grammatically similar in other respects – as reflected, let us say, in identical subjects and objects – might differ by membership in just such a covert category, and that the categorical difference was reflected by a clear difference in meaning. It has proved to be most revealing to describe grammatical items in terms of complexes of recurring grammatico-semantic features, and thus far it has turned out that these complexes that make up an item in one language are paralleled in various languages. Thus it seems that it may be possible in a very exact way to speak about correspondence of features of words. The verbs of English, French, and Russian are now being analyzed according to grammatical phenomena within each language. We find that certain contrasting classes of transitive verbs recur regularly between languages, as in Russian "slušat" – "slyšat", English "listen to" – "hear", that classes of verbs motivating a given type of subordination also recur between language, and that by treating types of transitivity and types of subordination as features, which may occur together and be combined with other features, it is possible to make explicit the notion of an item of one language corresponding to an item of another and to predict "correct" correspondents to items that have not yet appeared translated. Also, it seems to be possible in terms of such features to explain the difference that is felt between renderings of items between languages – when the translation is "as good as possible" but does not "catch the full essence of the original."

E. S. Klima

D. FRENCH GRAMMAR

A generative grammar of French that is based on the Left-to-Right model is being written. Progress can be estimated from three points of view.

First, as a general framework of a grammar of French, the program is near completion. Checking out of the program on the computer is underway. Grammar rules have been written for producing many different types of sentences and a number of different syntactic structures, such as simple, compound, and complex sentences, declarative, interrogative, and negative sentences, complementary infinitive clauses, relative clauses, indirect statements, variety of subjects and objects, passive voice, and synthetic and analytic tenses.

Second, as an analysis of specific problems of French syntax that apparently must be solved for the purposes of mechanical translation, a number of problems have been solved insofar as this particular program is concerned; that is, ways have been found by which the constructions in question can be generated correctly. These solutions have not always been fully satisfactory linguistically, but they are effective. For example, the interrogative pronouns qui, que, qu'est-ce qui, instead of being generated from a source labeled "Subject" or "Object," are generated from a source labeled "Sentence Modifier" because (a) they function as a sort of Interrogative Marker and in that sense modify the sentence, and (b) their position in the sentence is determined by their interrogative function rather than by their subject or object function – and position is a very important factor in a left-to-right grammar. Nevertheless, their function as subject or object is accounted for, and the program obviates such a problem as duplication of subject.

However, there is still a number of problems, and it is difficult to judge how much time will be required to solve them. It is expected that the program itself can be used for finding practical solutions to these problems so as to complete the grammar more rapidly. At the same time, more satisfactory linguistic solutions to the problems will be searched for because such solutions are more likely to be generally applicable.

Third, various facilities of COMIT have been tried out as tools for making the grammar more efficient. For instance, there is a system of using shelves (readily accessible storage places) to store information about each word as it is generated. This information can be used to specify subsequent constructions that are dependent on the earlier word in some way. The use of shelves should permit one to account for such dependency despite intervening structures, and, in fact, should also permit one to set the necessary limits (agreement of person, number, gender, etc.) within these nested structures, without confusing the separate constructions. The development of techniques of this sort should be useful to the other members of the group who are working on German or English.

A completed generative grammar will be directly applicable to the problem of mechanical translation of French, certainly for translation from English into French. As a basis for a recognition grammar, the left-to-right model is perhaps the most useful because in this French grammar, so many important decisions (which affect subsequent words or constructions) are made at the word level rather than at higher constituent levels. The result of such a procedure is that a number of rules are devised for analyzing the possible functions of a particular word as soon as it is recognized, without having the necessity for an initial scan of the entire sentence.

D. A. Dinneen

E. ENGLISH AND GERMAN GRAMMARS

Work continues on the problem of how to write grammars that are to be incorporated into translation programs. The main over-all purpose of this research has been to develop corresponding grammars of the two languages that are to be treated in the translation process – English and German.

Concretely, we first translated an English children's story into German, and thus obtained a parallel text. By preparing the translation ourselves, we were assured of a better insight into the relationship between the two versions. V. H. Yngve began writing an English grammar of the text, while I participated in a group (Rosemarie Sträussnigg, J. S. Bross, and myself) who were writing a German grammar of the same form. Both grammars are written in the COMIT programming language, and in stages of increasing complexity, so that each stage can be run on the machine and thus tested. Thus far, only the early stages of the English grammar have actually been run, but the first two German programs are also ready to be run on the machine. These are, at present, generative grammars – grammars that produce outputs from the grammatical specifiers contained in the programs. It is felt that generative grammars of this type must be developed to a certain degree of comprehensiveness before the problem of writing analyzing (recognition) grammars for the input language can be dealt with concretely. It is envisaged that these analyzing grammars may represent modifications of the generative grammars.

J. J. Viertel

F. SEMANTICS

Work has progressed on developing a method for using the techniques of modern symbolic logic to introduce a semantical interpretation (definition) of those linguistic entities that function in a purely structural capacity. This is a grammatical category that was not recognized until the advent of modern logic, and is still not recognized by

many linguists who deal with contrastive analysis of grammars. These linguistic entities I call "structural constants" because they form the basic structure of the symbolic formulations that express the general structure of actual linguistic sentences. By differentiating these expressly grammatical features of a natural language system from those that are denotative (nouns, adjectives, verbs, etc.) and by regarding the former as constants, the latter as variables, an explicit schematization of sentence structure can be obtained.

Part of the problem is that of recognizing the linguistic entities that function as structural constants. The method of demonstrating whether or not a specific entity belongs to this grammatical category is to coordinate-by-definition this linguistic entity to its related logical constant, which is a logical structural constant belonging to a specific artificial language system, such as the truth-functional calculus of predicates, or the calculus of probabilities, and then to test, by means of the derivation rules of that system, whether or not such a coordination can be justified empirically; that is, it must be demonstrated that the linguistic constant has all of the structural properties of the coordinated logical constant.

Since natural language systems are much richer than any given artificial language system, additions have to be introduced, in a logically consistent manner, into the symbolic systems. Thus additions that were never formulated symbolically before have been made to the schematic formulations, in order to account for grammatical devices existing in the English language system. An example will suffice. The word "either" can act as a restricted selector-operator ranging over two variables. Here, a special symbol defined by means of already defined logical constants was introduced into the symbolic notation.

By properly augmenting the symbolic systems and by properly coordinating the logical symbols to the structural constants of a natural language system, a semantical interpretation or definition of these important structural linguistic devices can be made in a rigorous and testable way. Since it is the rules for combining these structural constants that must be explicitly formulated in order to semantically equate whole sentences to one another so that a sentence-by-sentence translation can be achieved, this method, in which use is made of the discoveries revealed by the construction of symbolic language systems, helps us discover rules of language that have hitherto escaped detection.

A detailed analysis has been made for many words that function as structural constants, and the notation for expressing their meaning has been worked out: although, still, but, any, either, either-or, not-(either-or), neither-nor, ever, even, if-then are some examples. These words have been given a semantical interpretation, and methods are being developed for recognizing just what meaning changes occur when the structural environment differs.

With respect to the analysis of the tense structure, progress has been made along these lines: Instead of regarding an event as a point, I now consider an event as having the properties of a line, so that every event must have both length and direction along a one-dimensional axis. By using "overlap" as the fundamental notion, instead of the ordering-relations of "before" and "simultaneous with," many new relations can be schematized; thus, since events can overlap, they can begin before or end after another event, and so on. It is shown that an event can be regarded as a point when the arrangement of the linear events along the axis fulfills specified conditions; that is, the relation of order holding between events that are regarded as points is a special case. The properties of the durations are also taken into account so that a classification of verbs can be established that will be useful in determining the overlap ordering of the speech-event with respect to the sentence-event and reference-event.

This formalized time-ordering schema is a general one that is applicable to all languages. Since the tense structure of any natural language can be coordinated with the formalized schematic structure, a means is given for coordinating the tense structure of the natural language systems with one another. Such a coordination is one of the prerequisites of adequate mechanical translation.

Elinor Charney

G. EVALUATION

At the present time, there exists no theory to aid in the solution of the central problem of machine translation – the transfer of meaning. Linguists have concentrated on the study of phonology, morphology, and syntax, leaving semantics for the future. Communication engineers have studied the transmission of various sign systems representing natural languages. Logicians have studied the semantics of artificial languages. It would be reasonable to hope that these three areas of study, taken together, would cover the major aspects of translation, and that a proper merging of their results would provide a basis for a theory of translation. However, such unification has not yet been accomplished. Whether it can be accomplished, or whether important pieces are still missing is now an open question. The research program described here is oriented by the belief that although the three areas of investigation indicated above are concerned with essential aspects of translation, none of them penetrates the hard core of the problem and answers the questions: How does natural language serve as an instrument of human communication? What are the communicative functions of the various linguistic processes? What are the universals without which translation would be impossible?

As a modest first step, a procedure is being established to evaluate proposed grammars of English quantitatively from the point of view of recognition. A system for representing the syntactical structure of English sentences has been adopted tentatively,

and the sentences in a corpus of English newspaper text are being coded accordingly. This coding is carried out by human coders who make full use of context in order to assign a unique structure to each sentence. Then the corpus will be analyzed by machine, with the use of recognition routines based on one or another proposed grammar of English, and a quantitative comparison will be made between the machine analysis and the human analysis. The next step will depend on the results of the present investigation.

D. Lieberman

H. ARABIC TRANSLATION

A preliminary program for the mechanical translation of Arabic into English has been completed and correct computer output has been obtained. A second, and more extensive, program is well on the way to completion. A thorough discussion of the second program is being prepared. The first completed program is capable of translating approximately 8000 sentences of the type illustrated below. The second program will translate 10^{10} sentences; this we consider a very conservative estimate. The vocabulary is extremely limited because our main purpose, at the present time, is to develop syntactic and morphological techniques, rather than to develop a dictionary.

These programs are composed of two parts: developing an analytical grammar of the Arabic input sentences, and a generative grammar of the English output.

In the analytical section of the program the computer makes an analysis of the Arabic input which is sufficient to extract the necessary information to generate an equivalent English output sentence. The computer does this by describing the grammatical structure of the input sentences and attaching thereto instructions to be applied to the generation of the output sentence. Upon completion of the analysis of the input sentence, the instructions attached to the analysis are applied to the generative grammar of the output language, English. These instructions restrict the computer to the generation of that sentence which is equivalent to the Arabic input.

An illustration of the computer analysis (minus instructions for generation of the English) will illustrate the level of analysis which has been reached.

YVRF ALAWLAD BNT. /ya'rifu l ?awlaada bintun./
* A GIRL KNOWS THE BOYS.

The Computer

1. Identifies Y as a possible subjective prefix.
2. Searches for a subjective suffix, WN or N.
3. Fails to find it.
4. Tentatively identifies the Y as "third person masculine singular subjective prefix."
5. Identifies VRF as equivalent to the English "know."

6. Identifies AL as "the."
7. Fails to find AWLAD.
8. Identifies A--A- as a possible plural affix.
9. Identifies /WLD as "boy."
10. Concludes that /AWLAD is equivalent to English "boy (plural)."
11. Identifies BNT as "girl."
12. Knows that BNT is feminine.
13. Recognizes BNT as indefinite.
14. Recognizes BNT as either nominative or oblique. (Does this because of lack of suffix A, /-an/, which occurs with this class of word as the allomorph of the accusative case, the remaining case possibility.)
15. Knows that a feminine noun not immediately following a verb may be the subject of a masculine singular verb.
16. Identifies BNT as the subject of YVRF.
17. Identifies /ALAWLAD as the object of YVRF, the only remaining alternative.
18. Initiates generation of the English sentence.

A. C. Satterthwait

I. INTERLINGUA

I have recently been studying the relevance of Interlingua to mechanical translation. There are various advantages in using an artificial language as a source language. In comparison with natural languages, an artificial language possesses greater syntactical regularity and a more restricted vocabulary; thus the recognition problem is simplified. In comparison with other artificial languages, Interlingua is of particular interest in that it appears to be easier to learn and to read than most of them. It has been increasingly used for summaries and abstracts, especially by medical journals. It does, however, possess a number of features that could be regularized, and thus it could be made more useful as a source language without, at the same time, making it harder to learn or to read; for example, the idioms, the reflexive verbs, the system of particles, and an occasional polysemantic word. Most probably, all syntactic and semantic ambiguities can be resolved by modifying or eliminating those features of Interlingua which give rise to them.

Of the remaining problems, routines for word-order rearrangement can probably be devised on the basis of present knowledge, while the automatic correlation of nouns with verbs, pronouns with nouns, and so forth must probably await the perfection of more powerful techniques.

J. L. Darlington

J. SYNCATEGOREMATIC ADJECTIVES

What is the linguistic status of what philosophers call "syncategorematic adjectives"? These are, for example, the adjectives in the phrases "real imposter," "intellectual dwarf," and "the former Mrs. Simpson." They are peculiar in the following respect. Most adjectives appear in (at least) two characteristic positions: attributive, as in "the long war," and predicative, as in "the war was long," but the syncategorematic adjectives appear only attributively. The hypothesis that is now in favor is that they may result from transformations of adverbial expressions, such as "really an imposter," "intellectually a dwarf," and "formerly Mrs. Simpson."

W. K. Percival

K. THEORY OF TRANSLATION

Here, attention has been focused on the translation of a particular construction in German, the passive. The problem is to discover whether or not there is a limited number of ways of translating this construction. If it can be shown that the number is reasonably finite, this fact might be construed as an indication that the over-all problem of machine translation is tractable. If not, our chances of making progress in the area of translation theory would seem small.

W. K. Percival

L. COMIT

The COMIT system, an automatic computer programming system for mechanical translation, information retrieval, and general symbol manipulation research has been a joint effort of the Mechanical Translation Group of the Research Laboratory of Electronics and the Computation Center, M. I. T. The system has been converted for the IBM 709 computer and is now working well. Its availability is increasing the progress in linguistic and mechanical translation research. We are now preparing the system for distribution to other computer centers through the SHARE organization. This involves reassembling the programs and putting the finishing touches on the comprehensive manuals.

J. L. Bennett, Carol Bosche, F. Helwig,
Muriel Kannel, V. H. Yngve