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

Session 3: CURRENT RESEARCH

CURRENT RESEARCH ON AUTOMATIC TRANSLATION AT HARVARD UNIVERSITY[1] AND PREDICTIVE SYNTACTIC ANALYSIS

Anthony G. Oettinger and Murray E. Sherry
Harvard University

The problem of automatic language translation has been studied at the Harvard Computation Laboratory intermittently since 1950, and by a growing group since 1956. Although the entire problem has been kept in mind at all times, it is generally known that the bulk of our past effort has been concerned with the compilation and the operation of an automatic Russian-English dictionary intended to serve both as a component of an automatic translator, and as an interim approximation to a translator. This phase of research, supported by the National Science Foundation and the Rome Air Development Center of the United States Air Force, has now been completed, but a small-scale program of continuing quality control has been undertaken.

The Harvard automatic dictionary file contains at present over 30, 000 entries representing about 15, 000 distinct words or over 150, 000 distinct inflected forms, suitable for use in mathematics, electronics and allied fields. The methods of compilation and operation have been described in a series of papers and reports [1, 2, 3, 4, 5] and will appear in detail in the book Automatic Language Translation: Lexical and Technical Aspects to be published by the Harvard University Press in the Fall of 1960. Responsible investigators are welcome to avail themselves of the dictionary file in whole or in part and, if they wish, of the associated compiling, updating, and operating programs as well. We believe the dictionary system to be fundamentally sound, of great accuracy and reliability, readily adaptable to any technology, and potentially of very high efficiency and economy. Mr. Sherry's paper "Automatic Affix Interpretation and Reliability of the Harvard Automatic Dictionary" in Session 7 deals with the culminating phase of dictionary research, and presents data regarding the reliability and accuracy of the system.

1 This work has been supported in part by the National Science Foundation and by the Rome Air Development Center of the United States Air Force.

At the other end of the bridge leading from Russian to English, we are currently investigating a portion of the problem of synthesizing English sentences, namely the classification and inflection of English words. As might be expected, this problem is simpler than the analogous problems in Russian although it is by no means negligible on a large scale. The paper "Automatic English Inflection" presented by Mr. Foust in Session 5 presents our results to date in this area.

Our major concern at present is with methods of analyzing the syntactic structure of Russian sentences. Our work in this area is based on the technique of predictive analysis first proposed by Mrs. Ida Rhodes of the Applied Mathematics Division of the National Bureau of Standards [6, 7] with whom it has been our privilege to collaborate; a collaboration that has proved to be an unusually exacting, stimulating, and rewarding experience for us.

Our research on predictive analysis has two main aspects, the experimental and the theoretical. Mr. Sherry discusses the experimental aspect in this session, and Prof. Oettinger's paper "A New Theory of Translation and Its Applications" in Session 8 deals with the theoretical aspect. The remainder of this paper is concerned with the significance of Mrs. Rhodes' brilliant and fundamental ideas, and of our theoretical development of these ideas.

It must be strongly emphasized at the outset that no claim is made of any final solution of the problems of automatic translation. Any such claim would be, in mildest terms, premature. The practice and theory of predictive analysis, however, do reveal that syntactic structures have an hitherto unsuspected degree of simplicity, regularity, and universality, and that, up to a certain point, they yield themselves to correspondingly simple and elegant, yet powerful, methods of analysis.

On the experimental plane, this simplicity is reflected in the extraordinary simplicity and lucidity with which all details of an algorithm for predictive analysis may be described, without recourse to intricate flow charts. Predictive analysis algorithms have "natural" separability properties, reminiscent of the clean-cut separability of certain mathematical problems in such "natural" coordinate systems as those provided by appropriate sets of

eigen-vectors. The algorithms reduce to a set of simple subroutines of two classes, each class being so homogeneous that a standard program frame may serve for all members of a class, which differ among themselves only in detail. The load of programming and coding, whether directly in machine language or by means of such elementary compilers as SOAP or UNISAP, may therefore easily be distributed among several persons, without risking chaos. Eventually, a simple compiler could be designed to produce subroutines directly from grammatical specifications. The subroutines are also independent from one another to a high degree reminiscent of the desirable property of those power series to which terms may be added without recalculating all others; this feature facilitates not only debugging, but also the analysis of the effect of any combination of subroutines. The complete details of the structure of our version of the Rhodes predictive analyzer, and of its effect on augmented texts produced by the Harvard Automatic Dictionary, will be given in a report now in preparation.

Our theoretical work has its genesis in a simultaneous contemplation of the predictive analysis technique of Rhodes, of the syntax of Lukasiewicz's parenthesis-free notation as given by Burks, Warren, and Wright [8] , of Chomsky's phrase-structure model of sentence synthesis [9] of certain syntactic concepts analyzed by Wundheiler and Wundheiler [10] of an explanatory model of English sentence synthesis outlined by Yngve [11, 12] and of the sentence analysis theories of Bar-Hillel [13, 14] and Lambek [15].

We were led to the belief that the technique of predictive analysis which--as given by Rhodes and applied to a natural language such as Russian, has an empirical, approximative, and iterative characfeer--must have an exact theoretical counterpart over some suitable simple artificial languages. It turns out that such languages and such a theoretical counterpart do exist: Lukasiewicz's notation is one of the simplest languages for which a suitable theory can be developed. All the languages studied to date in this connection are representations of tree structures, and therefore appear to include all languages with phrase-structure grammars in the sense of Chomsky, although the limitations imposed by the nature of the representations are not yet fully clarified.

We have devised extremely simple algorithms for translating back and forth among the parenthesis-free notation and several forms of the conventional parenthetic notation. These algorithms have the following interesting properties:

(1) Internal storage consists essentially of a simple "push-down" store which may be regarded as the limiting case of Rhodes' prediction pool.

(2) The input formula is scanned in one direction only.

(3) Each symbol in the input formula is used once and only once and in sequence, eliminating the need either for storing the input formula in internal memory or else for rocking tapes back and forth [16, 17, 18, 19] .

(4) The amount of internal storage required is independent of the length of the input formula, and depends only on the depth of the deepest nest in the formula.

(5) The symbols of the output formula are generated in proper sequence and in one direction only. Insertions or rearrangments are never necessary.

(6) The symbols of the output formula are generated practically simultaneously with the scanning of the input symbols, so that the translation is completed almost as soon as the last input symbol is read.

(7) The algorithms are easily devised and represented [20] in such a way that one can prove that they will be successful if, and only if, the input formulas are well-formed. Such algorithms are therefore ideally fail-safe.

(8) For each algorithm one can prove a so-called $Д_M$-theorem that has significant implications regarding both certain aspects of the predictive syntactic analysis of natural languages and the design of artificial languages, with corresponding fail-safe and efficient translators for automatic programming.

We believe that the significance of these theoretical results lies in the following:

(1) They provide a theoretical model that explains at least one essential feature of the Rhodes predictive analysis system. Preliminary results obtained by Sherry suggest that the present model can easily be extended to account for the other important features of practical predictive analysis.

(2) Together with the practical realization of predictive translation algorithms, they complement the theory of Chomsky which is concerned almost exclusively with sentence synthesis, by providing for the first time both a theoretical model and an empirically verifiable method for sentence analysis, which are consistent, at the very least, with phrase structure. Both the theory and the practice represent significant advances beyond the work of Bar-Hillel and of Lambek. It is evident that applications of the method and the model need not be limited to Russian.

(3) They are in accord with much of the observations and theory of Yngve about English, and with certain results of Miller [21] in psychology.

(4) They suggest a fresh and potentially fruitful and elegant approach to the question of intermediate languages in automatic translation and to the comparative study of syntax.

(5) They provide a new avenue toward the design of elegant, efficient, and fail-safe automatic programming systems [22] .

Predictive Syntactic Analysis

The technique of predictive syntactic analysis was first proposed by Mrs. Ida Rhodes of the National Bureau of Standards in February, 1959 [6] and was presented by her at the International Conference on Information Processing in Paris in June, 1959. Since this powerful and fundamental method apparently has not yet been generally understood and accepted, we wish to present a simplified version of the method as we have come to understand it by working with it since September of 1959. Our syntactic analysis program is still experimental and has not been put into a final form.

Predictive analysis is based on the following concepts, presented here in simplified form:

(1) Alternative functions - The starting point of predictive analysis is the information about the functions of words that is obtainable from a dictionary. Since the lexical properties of words do not always define a unique function, a set of alternative functions must be considered. For example, стол has two alternative functions: nominative singular and accusative singular.

(2) Prediction pool - The program analyzes every word in a sentence by attempting to fulfill predictions about the grammatical

function of that word. The predictions are stored in a prediction pool which is operated as a "pushdown" store, that is, the last prediction entered into the pool is the first one tested for fulfillment.

(3) Prediction span indicator - A prediction span indicator is assigned to each prediction indicating how long the prediction is to be allowed to remain in the pool. Three such indicators are:

(a) 00 - The prediction must be satisfied by the next word in sequence or not at all.

(b) 01 - The prediction must be fulfilled during the analysis of the sentence

(c) 02 - This prediction may be fulfilled more than once in a single sentence and therefore must never be erased from the prediction pool

(4) Selected function - The selected function is that alternative function assigned to an analyzed word by the program.

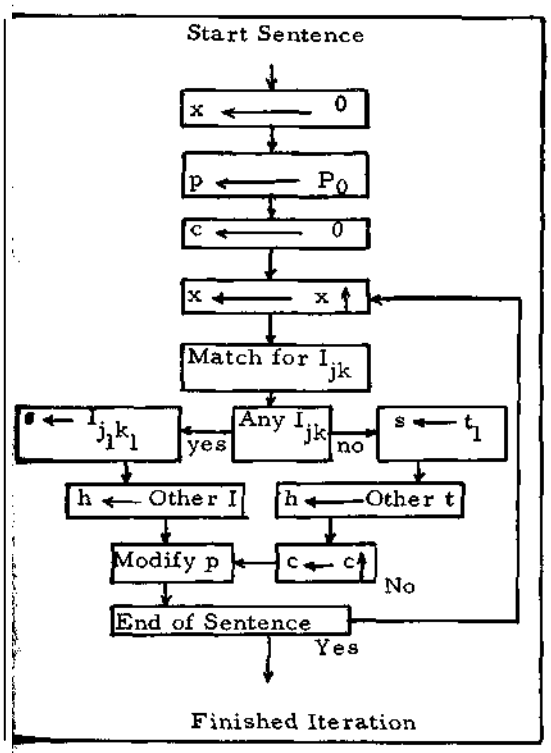
(5) Hindsight - During analysis, information that has to be stored, other than the selected function, is put into the hindsight. For example, if more than one function can be selected for a given word, all but the first, which is the selected function, are put into hindsight.

(6) Chain number - The chain number is an index that is incremented whenever the predictive analysis program cannot assign a selected function to a word based on the predictions stored in the prediction pool.

A simplified outline of the operation of the experimental syntactic analysis program is given in Figure 1. At the start of every sentence the program is initialized by inserting an initial set of predictions into the prediction pool and setting the chain number to zero. The alternative functions of the first word are compared with the predictions in the pool (Figure 2). Each prediction in the pool, in order, is compared with all the alternative functions of the word. The first alternative function that fulfills a prediction (the first "intersection") is accepted as the selected function. All other alternative functions that intersect with predictions are listed in the hindsight. The prediction pool is then updated. If there are no intersections, the first alternative function is arbitrarily accepted as the selected function and all other alternative functions are listed

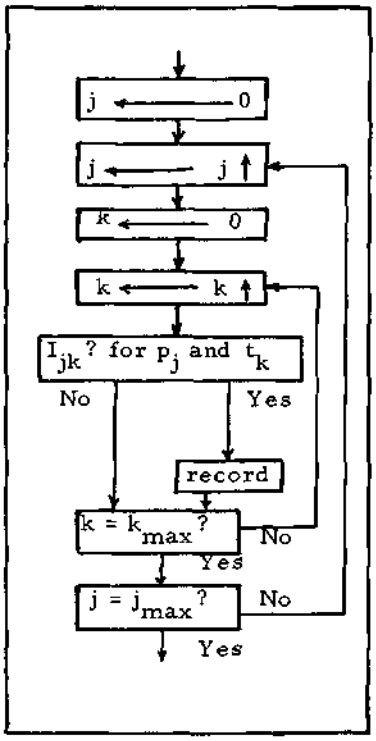
LIST OF SYMBOLS

x - current word
t - alternative functions for word x
p - predictions in prediction pool
s - selected function
h - hindsight
I_{jk} -"intersection" between p_j and t_k
j, k - indices
c - chain number



PREDICTIVE SYNTACTIC ANALYSIS

Figure 1



MATCH FOR I_{jk}

Figure 2

in the hindsight. This is an indication that there has been an error in analysis and the chain number is incremented. Various methods of using this error information to correct the analysis are possible. This entire process is repeated until the end of the sentence.

The following are among the significant properties of predictive analysis:

(1) It is the first approach which not only acknowledges that errors will occur during analysis but also provides a mechanism to detect and even to correct many of these. This is a firm step toward a fail-safe system, presently non-existent,

(2) We believe that necessary and probably sufficient conditions for a correct syntactic analysis of a sentence are that the chain number be zero and that there be no predictions remaining in the pool with a prediction span indicator of 01 after the sentence has been analyzed.

(3) In the event of an error, it is still possible to analyze nested structures, such as phrases and clauses, that follow the error.

Complete details of this approach will be given in a report now under preparation.

REFERENCES

[1] Giuliano, V. E., "An Experimental Study of Automatic Language Translation", Doctoral thesis, Harvard University (Mathematical Linguistics and Automatic Translation, Report No. NSF-1, January 1959).

[2,3] Mathematical Linguistics and Automatic Translation, Reports No. NSF-2 and NSF-3 (March and August 1959).

[4] Oettinger, A. G., Foust, W., Guiliano, V. E., Maggassy, K., and Matejka, L. "Linguistic and Machine Methods for Compiling and Updating the Harvard Automatic Dictionary", Preprints of of Papers for the International Conference on Scientific Information, National Academy of Sciences, National Research Council, Washington, D. C., Part V, pp. 137-160(1958).

[5] Giuliano, V. E., and Oettinger, A. G., "Research on Automatic Translation at the Harvard Computation Laboratory", Preprints of Papers for the International Conference on Information Processing, UNESCO, Paris (1959).

[6] Rhodes, I., "A New Approach to the Mechanical Translation of Russian", National Bureau of Standards, Washington, D. C., unpublished report (February 1959).

[7] Rhodes, I., "A New Approach to the Mechanical Syntactic Analysis of Russian", National Bureau of Standards, Washington, D. C., unpublished report (November 1959).

[8] Burks, A. W., Warren, D.W., and Wright, J.B., "An Analysis of a Logical Machine Using Parenthesis-Free Notation", MTAC, Vol. VIII, No. 46, pp. 53-7 (1954).

[9] Chomsky, N., Syntactic Structures, Mouton, ' s-Gravenhage (1957).

[10] Wundheiler, L., and Wundheiler, A., "Some Logical Concepts for Syntax", Chapter 13 in Locke and Booth, Machine Translation of Languages, Wiley, New York (1955).

[11] Yngve, V. H., "Left-to-Right Sentence Generation", draft manuscript (1959).

[12] Yngve, V. H., "A Model and a Hypothesis for Language Structure", draft manuscript (1959) (to appear in Proceedings of the American Philosophical Society).

[13] Bar-Hillel, Y., "A Quasi-Arithmetical Notation for Syntactic Description", Language, Vol. 29, No. 1, pp. 47-58 (1953).

[14] Bar-Hillel, Y., "Some Linguistic Obstacles to Machine Translation", Appendix II in Report on the State of Machine Translation in the United States and Great Britain, Technical Report No.1, Hebrew University, Jerusalem (1959).

[15] Lambek, J., "The Mathematics of Sentence Structure", American Mathematical Monthly, Vol. 65, No. 3, pp. 154-170 (1958).

[16] Rutishauser, H., "Automatische Rechenplanfertigung bei programmgesteurte Rechenmaschinen", Mitteilungen aus dem Institut für angewandte Mathematik, No. 3, Birkhauser, Basel (1952).

[17] Dartmouth Mathematics Project, "Symbolic Work on High Speed Computers", Project Report No. 4, Dartmouth, New Hampshire (June 1959).

[18] Kanner, J., "An Algebraic Translator", Communications of the the ACM, Vol. 2, No. 10, pp. 19-22 (1959)"

[19] Ingerman, P. Z., "A New Algorithm for Algebraic Translation", Preprints of Papers Presented at the 14th National Meeting of the ACM, pp. 22-1 and 22-2 (1959).

[20] Iverson, K. E., "The Description of Finite Sequential Processes", Theory of Switching, Report No. BL-23, Section III, Harvard Computation Laboratory, Cambridge, Massachusetts.

[21] Miller, G. A., "Human Memory and the Storage of Information", I. R. E. Transactions on Information Theory, Vol. IT-2, No. 3, pp. 129-137 (1956).

[22] Gorn, S., "On the Logical Design of Formal Mixed Languages", Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia (1959).