R E P O R T R E S U M E S

ED 020 503 AL 001 202

MACHINE TRANSLATION RESEARCH DURING THE PAST TWO YEARS.

BY- LEHMANN, W.P.

PUB DATE MAR 68

EDRS PRICE MF-$0.25 HC-$0.48 10P.

DESCRIPTORS- *MACHINE TRANSLATION, *LANGUAGE RESEARCH, *COMPUTATIONAL LINGUISTICS, APPLIED LINGUISTICS, DEEP STRUCTURE, SURFACE STRUCTURE, TRANSFORMATION GENERATIVE GRAMMAR,

THE AUTHOR RECOUNTS THE RISE IN IMPORTANCE OF MACHINE TRANSLATION, WHICH TOGETHER WITH LANGUAGE LEARNING AND TEACHING COMPRISE THE MAJOR FIELDS OF APPLIED LINGUISTICS. MUCH OF THE RECENT THEORETICAL WORK ON LANGUAGE DEALS WITH THE PROBLEM OF THE RELATIONSHIP BETWEEN THE SURFACE SYNTACTIC STRUCTURE OF LANGUAGE AND THE UNDERLYING STRUCTURE. THE CURRENT DEBATE IS BETWEEN LINGUISTS WHO SET UP SEVERAL SYNTACTIC LEVELS AND THOSE WHO SEE UNDERLYING (DEEP) STRUCTURE AS SEMANTIC. THE AUTHOR DESCRIBES THE STRATEGY BEING USED IN COMPUTATIONAL LINGUISTICS AS BEING OF TWO TYPES. IN THE FIRST APPROACH, A SELECTED THEORY OF LANGUAGE IS ASSUMED TO BE PRODUCTIVE AND USEFUL, AND EFFORTS ARE MADE TO SIMULATE IT. IF THE THEORY SELECTED REGARDS THE RELATIONSHIP BETWEEN SURFACE AND DEEP STRUCTURE AS TRANSFORMATIONAL, EFFORTS ARE MADE TO PROGRAM SIMULATED TRANSFORMS ON THE COMPUTER. THE OTHER STRATEGY, "OPERATIONAL," IS AN ATTEMPT TO HANDLE AS MUCH OF THE LANGUAGE AS POSSIBLE FROM SURFACE STRUCTURE. EXAMPLES OF THE RUSSIAN SYNTACTIC ANALYSIS PRODUCED AT THE IBM THOMAS J. WATSON RESEARCH CENTER IN 1967, AND OTHER WORK IN TRANSFORMATIONAL SYSTEMS ARE ALSO BRIEFLY DESCRIBED. THIS PAPER WAS DELIVERED AT THE SURVEY SEMINAR IN COMPUTATIONAL LINGUISTICS IN HONOLULU, MARCH 1968, UNDER THE AUSPICES OF THE UNITED STATES-JAPAN JOINT COMMITTEE ON SCIENTIFIC COOPERATION. (AMM)

Machine Translation Research during the past two years.

by W. P. Lehmann
University of Texas, Austin

In attempting a review of machine translation research during the past two years I look on my role as merely a prompter of discussion. For there are many here who could survey the field with greater authority than I. To mention only two, I point out that Dr Wada has a far deeper knowledge concerning the computational theory involved. And Dr Kuno, as shown by his fine essay: "Computer Analysis of Natural Languages" has more knowledge than I of the linguistic theory involved. It was a gracious gesture of two of the leading men in the field to entrust me with the honor of leading off the discussion.

We may look on machine translation as an applied form of linguistics. In the past there has been only one prominent form of applied linguistics, language teaching. Now suddenly there are several. Apart from work on mechanical translation we also see communications engineers studying language, to get under control the formidable problems involved when an increasing number of three billion men on earth want to give and receive information in other ways than over the back fence to their neighbors. All of us know of further incipient applications of linguistics. They result from the increasing interest among a rapidly increasing number of mankind to communicate more widely, more immediately and more directly than has been possible in the past.

When any application of any science is attempted, it may be carried out from several points of view. Since we have most experience concerning applied linguistics in the field of language teaching, I review briefly some of the activities of practitioners in that area. A hundred years ago the problem was simple. If you wanted your son or daughter to learn French, you simply hired a French governess. It's hard to beat a situation in which a skilled expert handles your individual problem. But engaging as the prospect may be, we could not continue this situation. On the one hand, the number of boys and girls who set out to learn French surpassed the number of young French ladies who could teach them. On the other hand, young

moulding of recalcitrant young subjects in a second language.--About forty years ago another approach toward second languages became prominent in the United States. A distinguished commission came to the conclusion that widespread instruction in foreign languages was hopeless. As a result, language teaching was virtually abolished in our schools. As a further result, there was a fair bit of anguish around 1940 when it became clear that other people used other languages than English.--The machine soon came to be involved in the problem, when it became apparent that the tape recorder, which was made available about 1947, could be useful in the teaching of foreign languages. Since that time language teachers have held various points of view. There were the enthusiasts who thought the machine might do the whole job. Confronting them were the negativists who booted the tape recorder out of the academy. In the middle were most language teachers, some of whom use the tape recorder as a minor adjunct, others of whom rely on it much more widely. Whatever the practice of individual teachers, in the twenty years since the tape recorder became available for language teaching there has been a tremendous change in the procedures used. Few think it will do all the work, or even that linguistics is the only discipline needed for skilled language teaching. But virtually no one disregards it. And there has been considerable research on the most useful procedures to apply in using machines to assist in the teaching of languages. By some approaches the tape recorder is virtually a simulated language teacher. By others, it is considered most useful for restricted supplementary applications.

The history of one application of linguistics may not be without its parallels to another. It seems to reflect a diversity of approaches to problems, which have also been reflected in lighter comments on the diversity of mankind. One set of these concerns the reactions of a number of individuals to a charge to write a treatise on the elephant. Of the reactions I'll recall only four. One of the resulting treatises was entitled: "Fifty dainty ways of preparing the elephant." Another, something like: "The elephant's contributions to his master." A third: "A study of selected capabilities of the elephant." And all of us may remember

best a fourth treatise, produced by an author who retired to an attic and contemplated the problem for several years; his title read: "Is there an elephant?" Now the only thing wrong with this part of the story is that such a treatise could scarcely have been produced by an individual. It could only have come out of a committee.

In this sketch I will pass over the activities concerning machine translation during the past two years which reflect the "Is there an elephant?" approach. Nor will I need to spend any time on the hearty outsider, generally an extroverted humanist, who asks whether the machine will be able to translate Homer or a haiku; he would fit a fifth treatise entitled: "Can the elephant make a watch?"

I will also spend little time on the "fifty dainty ways" approach. Machine translation is being carried on as you all know. The Air Force is replacing the Mark II system, which is in its eighth or ninth year of use, though originally intended only for five. Of possible comments here the only one I consider interesting is whether this protracted use of an obviously tentative system reflects the intractability of the problem or rather that of the specialists who might have been at work to produce a better system. Other installations in Europe and at Oak Ridge, Tennessee are using forms of the Georgetown system; one of the remarkable situations in the general field of translation is that scientists last year should have requested 300,000 words of translation with this system, although its availability is not widely known; nor are the procedures simple for getting Russian materials translated with it. Dr Hutton of Oak Ridge generously provided me with a copy of a recent translation. You may have seen others. I have never eaten elephant, and have no idea how dainty a dish might be made of it. But I suspect that it would be used only in fairly desperate need; and I regard the application of this system at Oak Ridge as evidence for a desperate need among scientists to secure some form of immediate translation, whatever the shortcomings. But my basic interest here is not the social context of machine translation; rather, the scientific work carried out on it.

Of the groups working on machine translation, only one I know of is contemplating

the use of a special-purpose computer. The Air Force, in replacing the Mark II system, is planning to use general-purpose computers. Since this is the widespread point of view, I will not deal with the computational procedures involved in using computers to manipulate language. Clearly the programming is complex. But it is being directed by the needs of linguists. Increasingly it is held that the basic problem in machine translation is our inadequate understanding of language. Accordingly my chief concern will be the attempts to deepen our understanding of language so that we may manipulate it computationally, including for the purpose of machine translation.

In the history of contemporary descriptive linguistics, which we may date from about a hundred years ago, the persistent aim has been to get from the surface manifestation of speech to language itself. This was the aim of Baudouin de Courtenay and Kruzewski, who directed their efforts at the sound system of language. They proposed the "deeper" unit phoneme. Surprisingly little effort has been devoted to the study of language. It should therefore surprise no one that their insight was not widely applied until about forty years ago. In the meantime Saussure had again pointed out how the surface structure may be inadequate, using as his example, sižlapră.

Besides recalling Saussure's approach, this example may illustrate one of the recurrent hazards of the linguistic profession: the adoption of troublesome patterns to point out difficulties of prime contemporary interest. I may recall some of them. Half a generation ago, when linguists were exploring the limits of phonological signalling an example cited was: "The sun's rays meet." vs. "The sons raise meat." Whether either sentence would ever be uttered is uncertain. But like Saussure's example, this putative pair of possible English sentences illustrates that we cannot distinguish the difference between some utterances on the basis of phonological criteria alone.--Somewhat later, linguists cited as a widespread example of the inadequacy of the surface structure: "Flying planes can be dangerous." This and similar examples were generated to illustrate that the same surface structure may

have more than one underlying syntactic interpretation. Other examples abound. One of the most recent ones I've seen is given in an article on the approach applied at Edinburgh: "The girl guides fish." From the surface we cannot determine whether "girl" is the subject of "guides," or whether it modifies "guides" which in turn is the subject of "fish."

Although the surface structure of such sentences is inadequate to determine the underlying structure, it is all we have. And all that the machine has and will have. It is scarcely surprising that much of the recent theoretical work on language deals with the problem of the relationship between the surface syntactic structure of language and the underlying structure. In this work, problematic sentences have been cited such as: "John is easy to please." and "John is eager to please." Somehow in the first sentence, speakers of English derive a meaning comparable to the underlying sentence: "Someone else pleases John." While from the second they derive a meaning comparable to: "John pleases someone else." Currently there is a vigorous debate on the solution to the problem. One group of linguists sets up several syntactic levels; another sees the underlying structure as semantic.

Computational linguists too are centrally concerned with the problem, and have been for some time. Disregarding various details, I would like to suggest that two types of strategy have been used and are being used in computational linguistics to deal with the problem. And it is in this framework that I would like to comprehend the work in machine translation over the past two years. By one strategy a selected theory of language is assumed to be productive and useful, and efforts are made to simulate it. If the theory one selects regards the relationship between the surface structure and the deep structure as transformational, efforts are made to program simulated transforms on the computer. By the other strategy, which I will call operational, an attempt is made to handle as much of the language as possible from the surface structure. After computer testing, the unaccounted portions will be studied further and computational procedures devised to handle them.

These are the two strategies I see in machine translation research of the past

two years. Obviously any simplification like this has shortcomings, some of which may be pointed out in the discussion. Moreover, neither strategy neglects the area of major concern of the other. The simulation strategy may not limit concern to selected processes, such as transformations but may also involve concern with surface structures. Conversely, the operational strategy not only seeks to handle as much as possible of the surface of a text, but it also proposes to deal with the unsolved facets, with the deep structures.

To characterize the approach, and the contributions, of proponents of the operational strategy I should like to take examples from work of a group which is not represented here, the Russian syntactic analysis produced by the group at the Thomas J. Watson Research Center and published in a report of October 1967. In its study, this group limited its materials severely. The grammar designed was based on a sample consisting of 160 Russian sentences taken from Pravda editorials and "a variety of references on Russian grammar" (68). The resultant grammar is described as "a relatively extensive preliminary set of grammar rules for surface structure recognition of Russian sentences." And the strategy is specified as regarding this grammar "as one stage in a cyclical process consisting of formulation, testing and review of grammar rules" (68).

Of the problems encountered in proceeding to the grammar I should like to point to the work in setting up a subclassification of Russian nouns. Twenty-five sub-classes are labeled, and interrelated in a tree structure. The classification is not unlike those produced by scholars attempting to provide a more complete description of Russian with no regard to the computer, or to machine translation. We may regard this result as a support of my earlier thesis that contemporary research towards machine translation and towards the goals of non-applied linguistics are parallel.

A further set of labels in the tree indicates the current status of our understanding of Russian. Many of the branches are labeled "other". Now we may always need to label clauses by some designation referring to residues; and "other" may be

as good a label for residues as any. But one has the impression that the classification is preliminary. This impression is corroborated by the undertaking of a study, commissioned to lexicographers in the Library of Congress, to obtain a finer subclassification of Russian nouns, in accordance with 86 criteria. The necessity of such a study illustrates forceably how inadequate are our descriptions, even of the surface structure of a language as thoroughly studied as Russian.--Another section of the report indicates how contemporary linguists must supplement the work of earlier scholars which is stored in the dictionaries and grammars that have been published. The report complains about the little information available on Russian adverbs (128). Details on the shortcomings of the handbooks are not pertinent here, though one that can be determined from the report is the procedure of classing entities morphologically rather than syntactically. The attempt at machine processing of Russian indicated quite forceably the inadequate descriptions of earlier grammars and dictionaries, through a poignant remark: "the bulk of the true adverbs were not in [the list of adverbs] at all, since they were handled ... as derived from adjective stems..." (128). Through these selected references to the report I should like to reflect the type of work that is going on in groups applying the operational strategy. I could cite other topics, such as the reflexive verbs and various types of Russian clauses that are under study. All of these are surface manifestations that must be understood regardless of the theoretical approach one holds to language. It is regrettable accordingly that work on machine translation has been reduced. The only support now provided is given by the Air Force. To be sure, any descriptive linguistic work under whatever support, will contribute ultimately to the computer manipulation of languages. But as past treatments of adverbs in grammars and monographs concerning Russian may indicate, without application such treatments may sweep portions of language under a rug. Probably anyone of us who has concerned himself with language, even a language studied only for purposes of research like Gothic, has been unhappy about the treatment of adverbs. But until we are forced to supply a better treatment, the old approach persists. As a parallel I may cite the acoustic analysis of language,

which remained somewhat sedate until possible applications supplied the means for the great contributions of the last decade. Similar attention to the syntax and lexicography, not only of Russian, but other major languages is urgently necessary.

Since my aim is to characterize the recent work on machine translation, I will not discuss other activities under an operational approach. A report on the procedures applied at Teddington was made available in Booth's collection of essays on machine translation. Similarly the work of Kulagina and Mel'chuk; although this has not been tested, it has avowedly practical aims. Booth's compilation also includes a report by Yngve, though it deals with contributions of his group up to 1965--a year before the promised bounds of my comments. Machine translation activities at Grenoble include semantic analysis, and in this way extend beyond attention to the surface structure. Grenoble also deserves mention here, because to my knowledge it is the only center outside Japan in which work is carried on in Japanese.

I conclude this statement on the operational strategy with a reference to Nida's review of Syntactic translation by Wayne Tosh, Language 42.851-854 (1966). While commending Dr Tosh on his detailed analysis of procedures based on surface structure, Nida suggests that machine translation might follow procedures similar to those he sees applied by man in translation; the proposed procedures call for "back-transformation of the source-language text to ... the deep structures." Further, for "transferring from the source to the target language at this deep level and restructuring of the message by forward transformation to the appropriate stylistic level in the target language" (854).

To carry out this procedure, we would need to be able to perform transformations with computers. The research which is being carried out to simulate such procedures I include under the simulation strategy. For any survey the cited paper of Dr Kuno's is of great importance.

As he has pointed out, one of the problems with this procedure is the shifting conception of language by transformational approaches. It is common knowledge that severe modifications have been suggested even for the model proposed by Chomsky in

1965. Nida in his review also indicated that "transformations of the source-language text [may need to be made to] a level much 'deeper' than most structural transformationalists now work" (854). Accordingly there are problems for the simulation strategy in addition to those encountered in attempting to simulate processes which have been proposed for achieving an understanding of how man uses language.

Such processes are exceedingly complex, and a description of procedures designed to simulate them seems too formidable to attempt in a brief oral presentation, even to a highly selected group. I may mention that transformational systems have been designed and tested at Mitre and at Brandeis, at the Thomas J. Watson Research Center and at Harvard. The Mitre and Brandeis systems apply so-called reverse transformation to syntactic descriptions obtained from analysis of the surface structure. By means of these reverse transformations, base structures are derived from the surface structure. These derived base structures, which correspond to real base structures or not, can be checked by comparison with base structures that can be generated during synthesis. In this way, some reduction of potential base structures secured from surface structures is possible. The entire procedure includes a number of grammars. Various kinds of transformations are possible. Analysis has been tested, and carried out in remarkably short time, in view of the huge number of surface trees that are produced for given sentences.

The Harvard and Watson Research Center procedures derive base structures without using reverse transformations. Devices that these use, such as virtual symbols, reflect advances that have been made in computer theory. With such devices and others that may be developed, the forbidding problems occasioned by transformational analysis may yield in time to computer manipulation.

I can only allude to other computational procedures under development, such as string analysis carried on by Harris and groups associated with him, or dependency analysis. But I may mention that at Grenoble dependency procedures are applied after syntactic analysis has been carried out, to deal with semantic structures. In this way flexibility is evident in current research, as Nida

recommended it should be.

Moreover, the two-pronged approach--on the base structure, on the surface structure--may contribute more rapidly to progress in machine translation studies than would only one similar strategy. The problems met in the study of language lead us to hope for such progress. On the whole, the field of machine translation has settled down during the past two years to the steady research that is necessary for progress in any area. The operational strategy is producing improved descriptions of individual languages, which in addition to their general interest will be essential for the initial steps in analysis and the final steps in synthesis of any language. The simulation strategy is providing techniques of general pertinence to any language manipulated computationally. Ventures are being made into little known areas, such as semantics. When these efforts are combined, progress in achieving computer manipulation for any selected languages will depend largely on the available effort and support.

Selected references:

Booth, A.D. (ed.) Machine Translation. North-Holland, Amsterdam, 1967.

Kuno, Susumu. Computer Analysis of Natural Languages. To appear in the Proceedings of the Symposium on Mathematical Aspects of Computer Science.

Plath, Warren J., Alexander Andreyewsky, Robert E. Strom, et al. Syntactic Analysis of the Russian Sentence. Technical Report No. RADC-TR-67-484. October, 1967.

W. P. Lehmann
The University of Texas
Austin, Texas 78712

Paper delivered at Survey Seminar in Computational Linguistics, Honolulu March 25-27, 1968 - auspices of US-Japan Joint Committee on Scientific Corporation (AHM, ERIC/CAL per Dr Lotz' Office)