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

Session 2; CURRENT RESEARCH

REPORT ON THE TEXAS PROJECT

Stanley N. Werbow
University of Texas

Since the machine translation project of the University of Texas is a new one and our project reports have had rather limited distribution, it may be well to review the background of our interest in MT. This interest was a linguistic one. That is, we at Texas have been occupied in varying degrees with research in German syntax, historical work with Early New High German, with analysis of modern written German, and also with the study of contemporary spoken German as reflected in the tape recordings of The German Language Archive at Münster, of which Texas is currently receiving a duplicate set.

Modern American linguistics in general has of late been turning its attention to syntax, and the last two Texas English Conferences have been devoted to the syntax of English. We have watched with great interest the work of the Massachusetts Institute of Technology group, which is based on the transformational grammar theory of Noam Chomsky, and of the Transformations and Discourse Analysis Project at the University of Pennsylvania. Professor Victor Yngve was good enough to visit us in Austin and discuss the aims and progress of his project, and he then kindly invited me to spend a summer in research with his group. His recent work of which we shall hear more here today, but especially as reflected in his article "A Model and a Hypothesis for Language Structure" [1], has been of great value to us.

Georgetown University too was very generous of its time and resources in informing us of the work there, and Professor Dostert, Dr. Brown, and Mr. Toma graciously visited Austin to advise us. In the fall of 1958 a weekly seminar was devoted to exploring the field of machine translation. The investigations of that first year were characterized on the one hand by a reliance on the incorporation of a maximum of information in the stored glossary and on the other hand by the desire for elaborate grammatical analysis routines to assure at least partial handling of material not available in the glossary by means of morphological recognition procedures.

It soon became clear that the writing of recognition routines and corresponding generation routines meant an enormously complex and voluminous program. Moreover, such a program would be so rigid that changes occasioned by continued research on an extended corpus, or by postediting large batches of machine translations, would require complicated and cumbersome revision by linguist and programmer or by linguist-programmer.

Our research at that time [2] was devoted to:

1. Sentence recognition and analysis. Essentially this involved only punctuation-recognition procedures.

2. Clause recognition and analysis. This used morphological and distributional information marking the finite verb, as well as punctuation handling.

3. Phrase recognition and analysis. Here verb forms and their distribution, prepositions and their uses, and analysis of the noun phrase were employed.

4. Word recognition and analysis. A major problem here is the compound word. Both intuition and experience made it clear that more effective procedures had to be devised for breaking down composite entities into units with higher recurrence probability; for, in German, freedom of composition is unlimited for all practical purposes, and any attempt to exhaust the entire stock of compounds by listing is doomed to failure from the outset.

5. Rearrangement of German word order at the phrase and clause level, a most troubling problem between German and English, is nonetheless probably of less complexity than many others.

The group was fully aware of the problems of lexical selection whether a minutely cross-coded inventory of the glossary is provided to produce fine distinctions and hence relatively polished translations, or whether a core-translation approach yields rough equivalents for a much less carefully prescreened subject matter.

These experiences in our first exploration of translation by machine strengthened our conviction that only a hierarchical approach to language could be fruitful for this application of linguistics.

In our first quarterly report [3] the three hierarchies were called graphemic, grammatical, and semological. Our work has been devoted largely to the grammatical hierarchy, which we treated on two levels, morphological and syntactical, with the internal breakdown: inflectional, derivational and phrasal, and clausal.

Our research team is headed by Winfred Lehmann and consists of linguistic analysis groups under the direction of Werner Winter and me and a programming group under Eugene Pendergraft. The work of the programming group, and in particular their recommendation that we apply the principles of the stochastic-phrase-structure grammar as suggested by R. J. Solomonoff in his Zator papers of April and October 1959 [4], has encouraged us to proceed with a grammatical encoding of large numbers of German sentences and their English equivalents in order to arrive at translation rules between German and English. Our coding objective is to have, for each word in a corpus, grammatical designations which are sufficient and necessary to convert the encoded sentence into a parenthesis-free notation or, in other words, back into the original graphemic form, Instead of writing the whole grammar in advance by impression and intuition, we shall build up a constituent-structure grammar from actual texts, using this to analyze further texts from which additions to the grammar will be derived. This process, though of vast extent because of the potentially unlimited corpus, should provide us with a finite grammar capable of generating an infinite set of sentences but producing only a finite set, as Yngve points out [5] .

Our contract from the Signal Corps brought with it certain requirements concerning the nature of the material to be translated and the conditions under which translation should be performed, including machines ultimately to be employed. In consultation with Signal Corps technical advisors, a basic policy decision was made early in the work, namely, that we would not neglect the statistical implications which derive from working with a corpus; that is, that while we would keep the total grammar of each of the languages involved as an ultimately desirable goal, we would be satisfied with—and indeed within the framework of our contract committed to—shaping an instrument for adequately translating those structures which occurred in a reasonably large, varied corpus of technical writing. With this decision, we consciously departed from the more ambitious goals set forth by Yngve (e.g., at the National Science Foundation conference in Washington in October) for the M. I. T. group.

With statistical probability came considerations of style, since we were now interested in the possible effect of style level, e.g., telegraphic style, on the complexity of the program required. Research by Werner Winter, reported to the Linguistic Society of America at its December meeting and covered in preliminary form in our First Quarterly Report [6] , indicated that genre does indeed affect the syntactic structure, as Winter demonstrated with the varying incidence of non-subject in first position, and with sentence length, in technical, journalistic, novelistic, and dramatic texts.

Our main efforts at present then are, on the language side, the encoding of the words and word components of German sentences, and on the machine side, the writing of programs for converting these coded sets of sentences into tabulations of grammatical structures. The tabulated grammar would include not only the permissible equivalents in English structure but also the frequencies of sentences and sentence components in the source and target languages. By means of these frequencies the most probable mode of generation of an input sentence is found and from that the rule for generating the target sentence by computing conditional probability. In this fashion both the lexicon, including word formation, and the sentence structure are part of the grammar. New lexicon material is added in sentence form so that the tabulation of the structures can take place. It is hoped that eventually the tabulation and probability computation for new lexicon material will become unnecessary. These programs are not quite ready, but we are encouraged that we are on the right track, at least on a right track.

REFERENCES

[l] V. Yngve, "A Model and Hypothesis for Language Structure" (work paper, unpublished).

[2] University of Texas Machine Translation Work Paper, January 31, 1959.

[3] University of Texas Machine Language Translation Study, First Quarterly Progress Report, 1 May 1959-31 July 1959, "On the Resolution of Subject-Object Ambiguity in German Texts".

[4] R. J. Solomonoff, A New Method for Discovering the Grammars of Phrase Structure Languages, (ZTB-124 Zator Company, Cambridge, Mass., April 1959. AFOSR-TN-110; ASTIA Document No. AD 219 390.]

The Mechanization of Linguistic Learning, [ZTB-125 Zator Company, Cambridge, Mass., April 1959. AFOSR-TN-59-246; ASTIA Document No. AD 212 226.]

A Progress Report on Machines to Learn to Translate Languages and Retrieve Information. [ZTB-134, Zator Company, Cambridge, Mass. , October 19597]

[5] V. Yngve, "A Model and Hypothesis for Language Structure", p. 13.

[6] University of Texas Machine Language Translation Study, First Quarterly Progress Report, 1 May 1959-31 July 1959, pp. 81-86.