APPROACHES TOWARD A MODEL OF LINGUISTIC BEHAVIOR

Osamu Fujimura

Paper presented at the Survey Seminar on Computational Linguistics, Honolulu, March, 1968.

N. Chomsky identifies grammar as a description of competence and distinguished it from a description of performance. He has been concerned mainly with theory of grammar in his discussion of linguistic theory.[1)] Katz and his coworkers discussed the semantic aspects of competence in the form of semantic interpretation processes and thus expanded the theory of grammar into an intergrated description of linguistic structure.[2)] The present paper intends to discuss performance in relation to the notion of competence and find some possible approaches in computational linguistics toward a description of linguistic behaviors.

Language is characterized as a multi-level structure, and the structures of sentences, or relations between structures of different sentences, appear in different forms in different levels. From an information theoretical point of view, a sentence as a structure is represented in a sequence of steps in the generative process of the phrase structure component that select rewriting rules and thus a particular type of sentence structure and finally particular lexical items to be used in particular syntactic positions. If we regard these selections as being specified by some code, the information content of the sentence can be represented by means of the "input codes" to the syntactic four-terminal. The logical design of the four-terminal (input-output device) without reference to the input scheme is nothing other than the grammar of the language, and its output is the sentence (represented in one of the grammatical levels). All linguistic information contained in the sentence can be given in principle in the form of selections of specific generative rules including lexical insertions. The selections of lexical items constitute most of the information content of the sentence, since in selecting a lexical item the number of candidates among which a particular choice is to be made is large. A quantitative evaluation of information requires knowledge of the statistical nature of the use of the selections.

The essential question in this point of view is in which of the descriptive levels of language the information is given. In other words, the crucial issue is what sort of information is linked with which kind of rules that allow, according to the grammar, freedom in choice. Occurrence of such input codes must be treated in the theory of performance, and the problem of their quantitative, in particular statistical measurements should be the central issue of computational linguistics. Facts about learning a language, either for a native child or a foreigner, must be accounted for by such input-output device as a self-organizing system. Creativity of language also may be interpreted as a random process at appropriate levels of the probabilistic description. The essential complexity of these linguistic phenomena will be explained partly in terms of the complexity of the multi-level structure of grammar, and partly by the complex relation between the environmental or internally created linguistic stimuli and the input codes. Difficulty of statistical treatment, within the framework of syntactic structures, arises when we do not know exactly how grammatical description interrelates the relevant grammatical levels, or more specifically what the transformational processes involved are.

In relating the generative process as a grammatical description to the behavior of perception (or production), only a generalized kind of analysis-by-synthesis model seems to promise solution.[3)] Applications like mechanical translation or automatic information retrieval naturally require due considerations of such probabilistic predictive processes. What may be meant by "consideration of meaning" in automatic processing would be treated in this way in various degrees of approximation, making the term clearer as the descriptive framework of grammar attains a better explanatory adequacy and the technical capability of handling a huge amount of cross-referential memory is achieved. The present computer systems, however, are not adequate for handling these problems in a straightforward mechanical fashion. We cannot expect any machine to discover the grammar relevant to the sample language that is given as data. It cannot analyze and recognize sample sentences, if these are fed into the mechanical system as surface forms. But this does not mean that we cannot study the statistical nature of language in a meaningful way. Our present interest is in the deep structure of sentence, since this is the formal representation that is most closely related to the meaning, and thus to the use of language.

In Aspects of the Theory of Syntax, Chomsky defines deep structure in a specialized way. According to this point of view, linguistic information given to a sentence is all included in its deep structure, and transformation has nothing to do in this regard. Whether this characterization of deep structure as against surface structure is solid is somewhat doubtful.[4] In any case, it seems to me that the Chomsky's deep structure contains different sorts of information, and a more powerful model may be claimed by separating some syntactic information from so-to-speak purely lexical information. The lexical information here includes the syntactic functions of the lexical items in the grammatical positions in use. Thus we may consider skeletal "core structures" of elementary (kernel) sentences.[3] The core structure pertains only to substantive items, and it identifies these items in the tree structure, thus relating these in some particular functions. It excludes elements like tense, determiner, negation, question, various sorts of emphasis including those associated with Kuroda's "attachment transformations."[5] The Japanese particles, for example, do not appear in this core structure except implicitly in the form of the abstract functions of the syntactic positions. These particles acquire their explicit forms in a later stage of surface structure formation, and the surface form would have its own aspect of information. As a first approximation in consideration of the statistical properties of lexical items, which is a major issue in the computational linguistics and is also a central topic of performance theory, we can exclude all these "surface aspects" of language and concentrate ourselves on problems of "langue."

There are many unsolved and pertinent problems, of course. In particular, there are problems as to which formatives in actual sentences are to be regarded as lexical items in the core structures. A surface sentence contains in general many core structures. Many problems arise in connection with appropriate analyses of adv.erbs and adverbial expressions.

In analyzing performance data, we can restrict ourselves, in starting, to treating obviously substantive items in given sample sentences in order to obtain approximate estimates of the use of these lexical items in the skeletal structure.

This will enable us to separate different problems. We will try to estimate separately the pattern of use of grammatical formatives. We will then proceed to evaluate interaction between these different elements and interpret the data in terms of the transformational structure of grammar.

We will have to resort to partially intuitive analyses of the surface form data in order to identify abstract core structures. A mechanical processing of such data on the abstract level will result in compilation and sorting of patterns of use of a selected range of lexical items. The range of interactions between items is highly limited and the relations are highly simplified in the core structure. Data of associative probabilities between items are meaningful when the syntactic relations in the core structures are determined. This kind of information, which may be represented in a lexon-taxon model,[3] must be present in a behavioral model of language use. Data extracted in this way can be utilized in constructing an analysis-by-synthesis model of sentence interpretation. Just the same model would be effective in testing the underlying grammatical hypothesis by random production of artificial sentences, if we supplement the core structure model with a limited but sufficient transformational component.

References

1) Chomsky, N.: *Aspects of the Theory of Syntax*, Cambridge, Mass.: M.I.T. Press (1965).

2) Katz, J. J. and Postal, P. M.: *An Integrated Theory of Linguistic Descriptions*, Cambridge, Mass.: M.I.T. Press (1965).

3) Fujimura, O.: "Some Remarks on the Analysis-by-Synthesis as a Model of Speech Perception," International Congress of Psychology, Seminar: "Speech Production and Perception," Leningrad (August, 1966).

4) Fujimura, O.: "Kotoba-no Kagaku(Science of Language)," *Shizen* (*Nature*) *21*, No's 2-12 (1966).

5) Kuroda, S. Y.: "Generative Grammatical Studies in the Japanese Language" (Ph. D. Thesis), M.I.T., Cambridge, Mass. (1965).