STRUCTURAL MODELS FOR LINGUISTIC AUTOMATION

W. P. Lehmann and E. D. Pendergraft

prepared for

National Science Foundation
Contract Grants NSF G-19277 and GN-54

LINGUISTICS RESEARCH CENTER
The University of Texas
Box 7247, University Station
Austin 12, Texas

LRC 62-WAA1 January 1963

TABLE OF CONTENTS

FOREWORD

This paper is the final draft of a contribution to the forthcoming book, Vistas in Information Handling. All publication rights are held by Information for Industry, Inc., 1000 Connecticut Avenue, N.W., Washington, D.C. Because the paper summarizes work accomplished under National Science Foundation Contract Grants NSF G-19277 and GN-54, the draft has been made available to the sponsor and to personnel of the Linguistics Research Center.

W. P. Lehmann
Director

1 LINGUISTICS AND THE COMPUTER

The formulation of explicit structural hypotheses is one of the striking developments of linguistics in the last decades. From the demand that linguistic analysis be based on form followed the study of formal properties and relations which were observable in individual languages. Since all known languages display underlying similarities, concern with form in linguistics has led to combined formulations by which the linguist might embrace the totality not only of a language but of language in general--that is, to structural hypotheses.

Two broad facets of research have accordingly developed in synchronic linguistics. One is toward a descriptive linguistics which establishes empirically testable generalizations about one or more languages. These generalizations display the structure of a language without seeking to explain it. The other is toward a theoretical linguistics which attempts to explain generalizations about structure by showing that they are logical consequences of more general hypotheses.

Although theorizing about language can be traced to antiquity, interest in basic linguistic research has recently been stimulated by provision of data from a tremendous variety of languages, and in the past few years by the availability of computers as means to test structural hypotheses. Data from a wide variety of languages has led to a broadening of theory

beyond that adequate to deal with the languages of Europe. Access to computers suggests the possibility of dealing for the first time with the multiple data of syntax and semantics as well as manipulating more comprehensively the data of phonology and morphology.[1] Descriptive and theoretical linguistics therefore promise to rely increasingly on the use of computers. Our purpose here is to examine the methodology employed in the science of linguistics, and to indicate some of its applications to linguistic automation.

For an awareness of the probable effects of linguistic automation on many scientific and practical activities is rapidly changing the supp[illegible]nd with it the character of basic research in linguistics. Accordingly we indicate how the conclusions of descriptive and theoretical linguistics are pertinent to such practical goals as mechanical translation, automatic abstracting, and information storage and retrieval.

Unfortunately experience has repeatedly demonstrated

---

[1]Since this is a programmatic statement, not a survey of research, it is scarcely necessary to list a great number of publications in support of these assertions. They may perhaps be supported most effectively by reference to the preprints of papers for the Ninth International Congress of Linguists, August 27-31, 1962, Cambridge, Mass. For the use of computers in linguistics see especially P. L. Garvin, "The impact of language data processing on linguistic analysis," pp. 331-337. The extent of our expanded knowledge of different languages may be illustrated by detailed problems discussed at the Congress, or by a glance at virtually any linguistic journal.

that these useful applications cannot be competently based on our present knowledge of linguistic structure.[2] Because language is so familiar we have understandably a tendency to confuse our prowess in using it with our lesser ability to explain it. Though we have come to understand that structural theories are a prerequisite to the construction of jet aircraft, we have not made the similar assumption that analogous theories are indispensable to linguistic automation. But without deep concern with theories of linguistic structure, we will not achieve an understanding of language, nor will we arrive at practical applications.

Our secondary purpose then will be to suggest how the computer may be used efficiently in research on the structure of language. For while the necessary understanding of linguistic structure might eventually be achieved through so-called practical experimentation, the testing of linguistic hypotheses on computers promises earlier and less costly success.

---

[2]Although numerous illustrations may be cited, we refer to our early progress reports which started from a linguistic basis inadequate for the operational requirements of translation. Our subsequent reports, produced under the direction of E. D. Pendergraft, entitled Machine Language Translation Study (Austin, 1959-) have been made widely available, obviating the necessity of explicit reporting here of progression in our theory. Our work was first supported under contract No. DA 36-039 SC 78911, under the U. S. Army Signal Research and Development Laboratory, Fort Monmouth, New Jersey; since September 1961 our more theoretical research has been supported by the National Science Foundation.

2 THEORIES OF LINGUISTIC STRUCTURE

To some extent structural theories have always been used in linguistic study, as we may indicate by reviewing briefly the sources of our current knowledge.

2.1 Greco-Roman Theory

The Greeks, dealing with language in the framework of human institutions, abstracted their structural hypotheses from man in relation to his surroundings. Dionysius Thrax, for example, defines grammar as the empería tõn parà poiētaĩs te kaì suggrapheũsin hōs epì tò polù legoménōn - the science of the standard writings by poets and prose writers.[3] He then enumerates the formal categories of his language and defines them, listing no paradigms nor much detail. The first type of word (part of speech), the noun, is defined as a class of words with cases, signifying concrete or abstract things, a concrete thing like 'stone', an abstract thing like 'education', referring to the general or particular, general like 'man, horse', particular like 'Socrates'. For the noun there are five subcategories: gender, type, composition, number, case. And so on. Applied to Greek the format was adequate, for it took

---

[3] For a brief introduction to western grammatical theory see R. H. Robins, Ancient and Mediaeval Grammatical Theory in Europe (London, 1951).

account of the structure of that language exhaustively, indicating all observable categories. But it was limited in scope, not proceeding as far as syntactic analysis.

The essentially realistic basis of Greek theory is probably most obvious still in grammatical terminology, as in the classing of certain nouns for their relationship to males, others for their relationship to females, and the remainder for their relationship to neither males nor females - Aristotle's metaksu 'between', the later oudeteron 'neither of the two'. This particular hypothesis was the target for the earliest theoretical criticism which has come down to us, in Aristophanes' Clouds, a part of which we cite from William Arrowsmith's translation (Ann Arbor, 1962) p. 53:

SOKRATES — I repeat: basket and Kleonymos are masculine in form and ending.

STREPSIADES — Kleonymos masculine? But he's feminine, Form and ending. Queer as they come.

Although the Greek hypotheses seem to us totally inadequate, especially as they were stated by the late Greek and Roman grammarians, we must note in their defence that grammar - the study of the tokens - was only part of the Greco-Roman theory, for it dealt with language in the wider context of philosophy.

The Sophists in particular concerned themselves with language in its relation to human nature and nurture; and following

the dictum of Protagoras that "Man is the measure of all things, of things that are, that they are, of things that are not, that they are not", they concentrated their efforts upon nurture. Because sophistic education emphasized the ability to speak well in furthering political virtue, inevitably the Sophists offered their rhetorical services to Greek citizens, analyzing in course the functioning of language. They also dealt widely with linguistic problems and included in their study the disciplines of syntactics, semantics, pragmatics and logic. Division of concern with language between grammarians on the one hand and philosophers on the other has persisted to the present.

Although they divided the study of language among competing specialists, the Greeks concerned themselves with the various orders of analysis which appear in the structural theories described below. Yet confusion between the orders, and insufficient breadth at especially the order of syntactics led to dissatisfaction with the Greco-Roman theory. Its shortcomings, or rather the shortcomings of its users, came to be particularly notable when the underlying hypotheses were applied inflexibly to other languages. For a language like English, the application was not disadvantageous as long as the users of the resultant grammars knew Greek and Latin; when these two languages were no longer mastered by educated men, details of the Greco-Roman theory seemed artificial and its rigid pattern was abandoned.

Still, the standard grammars in use in Western culture maintained the Greek viewpoint toward an 'empirical study' of language, though structural hypotheses were based on formal possibilities rather than only on the original categories, and though they became broader in scope, with special concern for spoken language. To illustrate the similarity of contemporary grammars with those of early Greco-Roman grammarians, we cite Harry Hoijer's structural sketch of Tonkawa, Linguistic Structures of Native America (New York, 1946) pp. 289-311. Like Dionysius Thrax, Hoijer presents only phonology and morphology. Moreover, he lists few paradigms, though he cites more forms than does Dionysius Thrax and he arranges some of the entities in charts.

As examples of differences, chiefly in terminology and independence from the categories of Greek and Latin, we may note that Hoijer sets up three classes of Tonkawa morphemes:

I. Themes (free and bound)

II. Affixes

III. Enclitics (bound forms with some restrictions)

Affixes may be of three types:

A. Transformative affixes; i.e., affixes by means of which a theme may be altered in function.

B. Verbal affixes; i.e., affixes which can be added only to verbs.

C. Noun and pronoun affixes.

Presumably with a sketch of this type the individual sounds and forms of a language could be generated, but as with Dionysius Thrax syntactics is excluded from the scope of the description, as is semantics.

In view of this venerable tradition, it is not surprising that the earliest experiments with mechanical translation employed the Greco-Roman theory, labelling words in the input language according to their categories, and seeking equivalent components in the output language. The depth of insight supporting these applications can be estimated by observing the resultant translations, some of which should be preserved along with the underlying analysis, if only for the interest they will have for historians of linguistics.

Clearly the Greco-Roman theory is not adequate to support translation. Hoijer's grammar, for instance, may be used to generate individual forms and sounds of Tonkawa, but certainly not sentences of that language; the latter aim would require a fuller structural theory including syntactics. The requirements of translation into Tonkawa, or out of it, are even more stringent, since translation involves necessarily a semantic relationship of 'equivalence' between linguistic entities. We conclude therefore that Greco-Roman theory provides only a scant beginning toward satisfying comprehensive structural requirements for linguistic automation.

2.2 Paninian Theory

As language again came to be intensively studied in the nineteenth century, dissatisfaction with traditional Greco-Roman grammars was heightened by the discovery of a separate linguistic tradition in India by which language was analyzed on a completely different basis. Instead of categories, the Indian tradition used rules to describe linguistic structure. The theory, furthermore, was highly developed, having produced the only relatively complete grammar then or now in existence, that of Sanskrit which is ascribed to Panini, a grammarian of the fourth century B. C.

The Paninian grammar has been widely praised, especially by scholars who have devoted years to its study, but to the casual observer it may seem baffling. To provide some basis for insight into it we cite the last of its several thousand rules, with a part of the comment of O. Böhtlingk, who provided the standard edition (Leipzig, 1887). The rule is simply:

a a.

In the taut form of most of the rules it indicates that the vowel a is actually a closer vowel than had been earlier assumed. For conciseness, rules throughout the grammar refer to only the short vowel of the short : long pairs, e.g., i for i : ī, u for u : ū and so on. Short a similarly was

treated as the paired member of $\bar{a}$, by a convention that reduced considerably the number of rules. But after all Sanskrit forms had been generated, at the very end of the grammar, the fictitious short open a is replaced through a single rule by the form which occurs in the language. Such a magnificent conclusion is part of the evidence that Paninian structural theory is relatively complete, that Panini's grammar of Sanskrit is an excellent example of the type, and that the theory would repay application to other languages.

But, in spite of the homage paid it by every elementary text on linguistics, the Paninian structural theory has not been fully verified. It could be tested, and the praise of its admirers perhaps confirmed, by programming Panini's grammar on a computer. Since Sanskrit has fallen into desuetude, such an effort will probably have to wait for the development of a Paninian grammar for a contemporary language.

We may forecast provisionally, however, that unless extended beyond the principles which have come down to us in the Indian tradition, the Paninian theory would scarcely be an improvement over the Greco-Roman for such purposes as translation, since again it deals with the generation of individual forms. The theory interests us primarily because it introduced the concept of a grammar consisting of rules. As we shall see, recent structural theories have in a sense combined these two ancient traditions by using rules to define the membership of linguistic categories.

2.3 Semiotic Theory

A behavioral frame of reference has come into use in linguistics, especially since the publication of Bloomfield's Language. Yet the same positivistic temper which introduced behaviorism to study of language so narrowed the subject matter of linguistic analysis that we must look elsewhere for more comprehensive theories to underlie linguistic automation. This broader framework has emerged from the tradition sustained through the Greek Sophists and the Hellenistic philosophies as a whole, the scientia sermocinalis of medieval Europe, diverging among the formalists following Leibnitz, British empiricists, and American pragmatists, and culminating finally in the writings of Peirce, Mead and Morris within the field of semiotic.

Semiotic, the study of sign-processes, has in particular begun to describe the complex and elaborate sign-behavior found in human speech and writing, though its primary contribution to date may well be the development of fundamental orientations and terminology for such an enterprise. In Morris' "Foundations of the Theory of Signs" (Chicago, 1938), for example, the sign-process itself is called semiosis. It is regarded as "involving three (or four) factors: that which acts as a sign, that which the sign refers to, and that effect on some interpreter in virtue of which the thing in question is a sign to that interpreter. These three components in semiosis may be called,

respectively, the sign vehicle, the designatum, and the interpretant; the interpreter may be included as a fourth factor." To differentiate between signification and denotation he says: "Where what is referred to actually exists as referred to the object of reference is a detonatum." The term designatum is replaced by significatum in later writings to further emphasize this distinction.

Three subdisciplines within semiotic are then distinguishable in terms of dyadic relations among these correlates of semiosis. Syntactics is characterized by Morris as concerned with "the formal relation of one sign to another", semantics with "the relations of signs to the objects to which the signs are applicable", and pragmatics with "the relation of signs to interpreters." He observes that the current tendency is toward specialized research in syntactics, semantics and pragmatics in consequence of attempts to systematize the extensive doctrines which have developed around the earlier viewpoints of formalists, empiricists and pragmatists, respectively. By providing a relatively simple theoretical framework and a common language with which to record results of sign analysis, semiotic has effectively challenged the role formerly claimed by logic as the organon of science. It may thus unify the flood of information produced by specialists in the study of sign-processes, about whose proliferation Morris remarks: "The army of investigators includes linguists, logicians, philosophers, psychologists,

biologists, anthropologists, psychopathologists, aestheticians, and sociologists."

Linguistics may accordingly be subsumed under semiotic as the study of sign-processes involving language signs. The theory of language signs is elaborated further in Signs, Language and Behavior (New York, 1946), where among other topics Morris outlines the principal modes of signifying and types of discourse which have evolved through specialization of language to various purposes. These details, though fragmentary, are sufficient to illustrate that the extremely complicated sign complex of a spoken or written language contains many modes of signifying and serves a vast variety of purposes.

Having chosen a behavioral orientation, therefore, descriptive linguistics now faces the enormous task of unravelling these modes and purposes in various languages, an undertaking of such magnitude that it could hardly be attempted without the data processing capabilities of modern computers.

This is nevertheless the task before us if we are to understand the structure of language with sufficient precision to make linguistic automation a reality; for the computer, as yet unable to produce its own description of English or any other language, has to rely entirely upon ours. More exactly, our formational description of any language must provide the computer with the information by which it would recognize or produce extant expressions of that language; our transformational

description on the other hand would provide it with the information by which to recognize or produce the extensions of extant relations among these expressions. Formational description is therefore a fundamental requirement for linguistic automation in that it is presupposed by transformational description; the latter in turn is presupposed by any process which utilizes relations among the expressions, that is, processes of inference, rhyming, paraphrasing, and so on. Transformational description for translation must additionally provide interlingual rather than monolingual information --that by which the computer would recognize or produce the extensions of extant relations among the expressions of two or more languages.

These are as a consequence the basic data for linguistic automation; until reliable formational and transformational information is available, applications will continue to be based on approximations which seriously limit their usefulness. Especially damaging is the lack of formational information, since the inability of current algorithms to accurately recognize and produce language expressions distorts the inputs and outputs of all potentially useful linguistic information processing applications.

How then are these vital language data to be recorded? To this question semiotic theory does not give a specific answer, nor does it explain precisely how linguistic algorithms would take account of sign vehicles, significata or interpretants.

It suggests instead general criteria for classifying formational or transformational data according as they are based on syntactical, semantical or pragmatical relations. The classification is assumed to be hierarchical, that is, by virtue of the basic relations among correlates of semiosis, pragmatics presupposes semantics as semantics does syntactics. Consequently the valuable contribution of semiotic theory has been to indicate, as a guide to more explicit theories, some of the fundamental levels of symbolization which occur in language.

2.4 Formalized Theory

In the last century formalization has come to be generally regarded as the desideratum of scientific explanation, since it offers a way to gain maximal clarity and explicitness in the presentation of theories. With Leibnitz proposal for an ideographic universal language the scientific community began increasingly to bypass the ambiguity of natural languages by constructing formalized languages wherein each symbol had a single meaning. As this tradition developed in logic through such early efforts as those of Lambert, DeMorgan and Boole, and then in mathematics through proponents like Frege, Peano and Whitehead, the advantages of formalized theory spread progressively to other fields of study greatly influencing the development of contemporary science.

That the methodological principle exemplified by Panini's rules took so long to germinate in linguistics was perhaps symptomatic of a youthful science still preoccupied with observation and classification. It was very likely also the result of a hazard peculiar to the application of formalization in linguistics, an appreciation of the great difficulties which would be encountered in any attempt to formalize a natural language by describing its rules of formation and transformation. About such an attempt even an enthusiastic formalist like Carnap has contributed a pessimistic opinion; he says in his introduction to The Logical Syntax of Languages (London, 1937) p. 2:

"In consequence of the unsystematic and logically imperfect structure of the natural word-languages (such as German or Latin), the statement of their formal rules of formation and transformation would be so complicated that it would hardly be feasible in practice."

This discouraging prediction should cause us to consider carefully the potential uses of formalized theory in linguistics. Above all we must distinguish the prescriptive and descriptive uses of formalization.

By prescriptive formalization we mean those uses where a formalized language is actually being constructed for some specialized purpose by a procedure such as the following one described by Church in his Introduction to Mathematical Logic (Princeton, 1956) p. 48:

". . .we begin by setting up, in abstraction from all considerations of meaning, the purely formal part of the language, so obtaining an uninterpreted calculus or logistic system. In detail this is done as follows.

"The vocabulary of the language is specified by listing the single symbols which are to be used. These are called the primitive symbols of the language, and are to be regarded as indivisible in the double sense that (A) in setting up the language no use is made of any division of them into parts and (B) any finite linear sequence of primitive symbols can be regarded in only one way as such a sequence of primitive symbols. A finite linear sequence of primitive symbols is called a formula. And among the formulas, rules are given by which certain ones are designated as well-formed formulas (with the intention, roughly speaking, that only the well-formed formulas are to be regarded as being genuinely expressions of the language). Then certain among the well-formed formulas are laid down as axioms. And finally (primitive) rules of inference (or rules of procedure) are laid down, rules according to which, from appropriate well-formed formulas as premisses, a well-formed formula is immediately inferred as conclusion."

These features included by Church in his logistic system correspond to those which belong to the syntactical description of natural languages. The complete description is then obtained by

adding semantical rules which provide the calculus with an interpretation through specification of the explicit objects or situations denoted by well-formed formulas. In practice this is usually accomplished by describing the denotata of certain primitive symbols. The calculus may have many interpretations--indeed its usefulness for scientific explanation may depend upon this property--though there may be a principal or standard interpretation to guide our intuition. This convenience is not an essential part of formalization and it may in fact be detrimental. For example it seems probable that the above terminology, which has its origin in the logical and mathematical interpretations of formalized languages, have been prejudicial against the use of formalization in linguistics.

It should be noted that the entire apparatus of proof is supplied by the syntactical description of the formalized language. As I. M. Bochenski points out in his recent book, A History of Formal Logic (Notre Dame, 1961), Bolzano was a noteworthy precursor of this important discovery, but it was Frege who developed the notions of proof-theory with greater clarity than ever before. Hilbert led the way to applications of proof-theory in mathematics and Łukasiewicz in logic. The notion of a proof is explained by Church as follows, p.49:

"A finite sequence of one or more well-formed formulas is called a proof if each of the well-formed formulas in the

sequence is an axiom or is immediately inferred from preceding well-formed formulas in the sequence by means of one of the rules of inference. A proof is called a proof of the last well-formed formula in the sequence, and theorems of the logistic system are those well-formed formulas of which proofs exist."

Since the formalization of pragmatical relations is still wrapped in controversy, the part played by interpretants of formalized languages is not well understood. Evidently however formalization involves certain operational commitments, called by Church, p.50: "requirements of effectiveness as follows: (I) the specification of the primitive symbols shall be effective in the sense that there is a method by which, whenever a symbol is given, it can always be determined effectively whether or not it is one of the primitive symbols; (II) the definition of a well-formed formula shall be effective in the sense that there is a method by which, whenever a formula is given, it can always be determined effectively whether or not it is well-formed; (III) the specifications of the axioms shall be effective in the sense that there is a method by which, whenever a well-formed formula is given, it can always be determined effectively whether or not it is one of the axioms; (IV) the rules of inference, taken together, shall be effective in the strong sense that there is a method by which, whenever a proposed immediate inference is given of one well-formed formula as conclusion from others as premisses, it can always be determined effectively whether or

not this proposed immediate inference is in accordance with the rules of inference."

The analogy of these requirements to similar concerns of descriptive linguistics should be obvious. Without elaboration we observe merely that they are requirements which must be satisfied by the language being used to describe the formalized language--not by the formalized language itself. The former is often called the meta-language, the latter the object language. The meta-to-object language relationship is a relative one characterized by the fact that denotata of the meta-language are features of the object language.

Prescriptive formalization accordingly focuses our attention upon the object language being designed and constructed as a specialized vehicle for scientific explanation.

The term descriptive formalization we have reserved for those contrasting uses where our purpose is to describe the structure of some one of the several thousand languages which already exist. Clearly in this case our interest is centered upon the meta-language as the vehicle for scientific explanation.

The illusive distinction between prescription and description was resolved in logic and mathematics only after protracted debate. Because at base it merely reflects the specialized interests of theoretical and experimental science which have already emerged in other fields, the consequences of

its approaching resolution in linguistics can be predicted with reasonable confidence. We will argue therefore that, although the contributions of formalized theory in linguistics are still undecided, they may ultimately be expected to furnish a sound foundation for linguistic automation.

3 STRUCTURAL MODELS

What direction will the solution to these problems take? Although it is perhaps too soon for explicit answers, some aspects of future theories are already apparent. In conclusion, we propose merely to indicate these new developments in linguistics and to characterize any theory resulting from them by the term structural model.

Unfortunately the term model has been used in scientific discourse in at least three distinct senses. In mathematics, for instance, a theory may be axiomatized by defining a set-theoretical predicate; any entity which satisfies that predicate may be called a model for the theory. In economics, what the mathematics calls a model may be called a structure, the model for a theory then being the collection of all its structures. A third sense, and the one here intended, has had widespread use in empirical science, where theory often has the connotation of inexactness, while model or mathematical model connotes more precise statement.

3.1 Descriptive Formalization

When mechanical translation was undertaken about a decade ago, linguistics had proceeded a long way toward use of descriptive formalization. A conspicuous start had been made by Bloomfield in his postulates, published in 1926. These

constituted a type of declaration of independence for the Linguistic Society of America. Yet the theories we find in writings of Bloomfield and his immediate circle were severely limited to the phonological and morphological levels of language, in spite of the association of Bloomfield and Morris in a joint publication series dealing with the foundations of science. Only since Bloomfield's death has descriptive formalization been carried to the study of syntax, notably by Harris and his students, for example by Chomsky, and later by Pike.

The limitations on the study of language adopted by linguists were not admitted in the more general study of signs. For Morris, as we have indicated above, the work of Bloomfield and his successors constituted simply one part of the study of language--syntactics--with semantics and pragmatics left untouched. Investigated by others, primarily logicians, these have developed their own methodologics, often forbidding realms of terminology for linguists. Yet they have prepared the way to concern with language in its totality.

But semantic and pragmatic studies have thus far been pursued primarily by means of prescriptive formalization--through formalized object languages designed either for philosophical analysis of theoretical concepts or for the approximation of natural languages by ones constructed more simply. Though research in mechanical translation has recently been

pointed toward the use of descriptive formalization in semantics, the work is extremely tentative. Four theories have been seriously proposed as a basis for automatic semantic analysis, for example, by Ceccato in Italy, Masterman in England, and Lamb and Pendergraft in the United States.

Recent concern with descriptive formalization has become very prominent, as in the work of Tesniére, and especially among investigators who are attempting to automate linguistic processes. Despite its self-imposed limitations, therefore, linguistics is recognizing the methods of functioning of language--that it is a system of signs with a hierarchical structure--and that for an adequate understanding of language it is essential to make use of descriptive formalization.

3.2 Formalized Meta-languages

Because of the difficulties presented by these more comprehensive theories, it seems plausible that linguistics will not only come to rely on descriptive formalization, but will do so by means of formalized meta-languages. This trend is even now in progress as a result of earlier work by Hjelmslev and Uldall in Europe, by Bar-Hillel in Israel, and in this country notably by Chomsky, Harris, Solomonoff and Yngve. There are several reasons to expect its further development.

The great practical difficulties encountered in descriptive formalization are essentially a consequence of the

empirical constraints within which the linguist has chosen to work. In making his description, he does not attempt to simplify the object language nor to reduce its ambiguity; his motive is to account fully for the speaker's ability to understand and produce utterances of the language even though some of them may be vague, ambiguous or wholly untenable in experience. His explanation will of necessity be complex and, by comparison to the requirements of prescriptive formalization, will place much greater demands on the meta-language used for his description.

If the meta-language is not formalized, these demands may be made on a language which is itself vague or ambiguous. Vagueness will then be described vaguely, ambiguity ambiguously, and so on. As a consequence the requirements of effectiveness in descriptive formalization may fail to be satisfied.

One of the pressing functions of linguistics today is accordingly to scrutinize more carefully the complex meta-languages which have been developed over long periods of concern with the description of language. In the past this scrutiny has proceeded chiefly from the application of an apparently adequate meta-language to additional natural languages. Phonological theory of today, for example, was developed largely through application of the widely used Greco-Roman meta-language to the indigenous languages of Africa, those of the Caucasus and those of the Americas.

This method has obviously contributed greatly to the explication of structural theories and is still productive. Yet explication of structural concepts through prescriptive formalization of descriptive meta-languages is clearly a more powerful technique. For by it structural concepts are not merely brought under scrutiny because of the inadequacy of a time-honored meta-language to handle features of a new language, but rather because of a need to develop comprehensive theories.

Theoretical linguistics has therefore emerged out of concern with descriptive meta-languages, much as descriptive linguistics grew out of concern with natural languages. However, its method is prescriptive rather than descriptive formalization.

3.3 Generalization

When language has been examined on all levels of its functioning, our understanding of it and consequently our manipulation of it will be considerably improved. But it is already evident from the scope of required procedures that these tasks will not be accomplished by the traditional type of humanistic study--a well-trained expert at a desk bounded by carefully selected reference materials.

The computer can be of assistance to the linguist in descriptive formalization through its capability to process language data. Indeed, as mechanical translation research progressed, increasing attention was given to the conclusion that

such information processing applications must be continuously supported by language data processing facilities which seek to improve the quality of structural description. For it would be uneconomical, if not impossible to describe natural languages fully before putting their data to use.

The feasibility of language data processing applications depends, in turn, upon their generality. Unless programming costs can be distributed to many languages, the linguist can be provided only inconsequential aids. Hence, to be economical, language data processing should be based on generalized linguistic processes.

Having proceeded to the formalization of descriptive meta-languages, linguistics may now be approaching a solution to this important problem of generalization--by answering the persistent query: "What is a rule?"

In the formalism being developed by the authors a rule may be viewed as an axiom of the descriptive meta-language. The process of linguistic synthesis therefore becomes one of deductive inference within the meta-language, and linguistic analysis one of inductive inference. Proofs and hierarchics of proofs then satisfy the requirements of effectiveness for the formational description.

This conclusion has led to the development of generalized algorithms for linguistic analysis and synthesis, and for translation. For all the information required for computer

programming may be derived from the rules of formation and transformation of the meta-language. This may be regarded as the operational interpretation of the underlying structural theory.

After he has specified the formational and transformational rules of his meta-language, the linguist may complete his formalized language by supplying its appropriate axioms and semantical rules, as his description of some natural language. This may be regarded as the linguistic interpretation of the underlying general theory by which it is specialized to individual languages.

Structural models having these properties will make possible the accumulation of descriptions of many languages in language data processing systems. Because such models admit to many different operational interpretations, linguistic information processing systems will utilize these same data for the various practical purposes mentioned above. Others will be suggested by the more adequate knowledge of language preceding linguistic automation.