LINGUISTICS AND AUTOMATED LANGUAGE PROCESSING[1]

0.1 This paper is concerned with natural language, computers, and two groups of people interested in natural language: linguists, and persons engaged in computer processing of natural language data. There is some intersection of the latter sets, but the intersection is quite small relative to the size of the sets themselves and is thus inadequate to provide linguists with a proper perspective on automated language processing, or computer scientists with a proper perspective on linguistics.

Although both groups of persons have a mutual interest in natural language, their conceptualizations of the nature of language and their approaches to processing language data are very different. To present a somewhat oversimplified view of these differences: linguists tend to be theory-oriented--they are concerned with interesting but sometimes quite esoteric problems, counter-examples, and the infinite set of sentences of competence; on the other hand, persons engaged in automated language processing tend to be data-oriented, and are concerned with statistical significance and with some finite subset of the sentences of performance. The question therefore arises as to whether these different perspectives are to be interpreted as incompatible or complementary, and if complementary, whether some research concept might provide the means for a unified approach to analysis of natural language.

In this paper, Section 1 deals with the perspective of linguists on automated language processing and computer scientists on linguistics; Section 2 discusses their respective concepts of natural language and

---

[1]I am indebted to Paul Garvin for his valuable comments on this paper.

and their approaches to analysis of natural language, and explores the questions raised above; Section 3 presents some concluding remarks.

1.1 It is appropriate to begin this discussion with a brief inquiry into the sources of the common focus of linguists and computer scientists on natural language. The interest of linguists in natural language is given by definition; the interest of non-linguistically oriented computer specialists in natural language derives not from a concern with language per se, but from the function of language as the primary vehicle for communicating information in human society. Whether or not one accepts the idea of the so-called "information explosion," the processing of natural language text is an important challenge for both linguistics and computer science. The sheer volume of natural language information is taxing manual systems to the point where most organizations which engage in large scale information processing are turning to automation of operations on natural language text. While most linguists are speculating on the theory of language, computer scientists with little or no linguistic background are attempting to construct systems for analyzing the content of natural language materials. Obviously, linguists should become involved in this development, but to date, few linguists have been motivated to participate. It appears that there are two basic reasons for the current lack of involvement.

1.1.1 In the first place, among linguists there is little appreciation of the fact that in essence, all processing of natural language information--whether scientific, technical, or literary--is a linguistic problem. Basically, the processing of natural language information for indexing, abstracting, fact retrieval, translation, or any other purpose requires an analysis of the content of the text and the representation of it in some standard form. Ideally, content analysis consists in determining the concepts present in the material and the interrelations existing among those concepts; the former is based on some form of semantic analysis, and the letter implies syntactic analysis, although the two forms of analysis are interdependent to a considerable degree.

The concepts and relations which have been identified are then translated into a set of canonic sentences representing the content of the document. From this representation of the content of the document, all document surrogates--such as strings of index terms or thesaurus groups, abstracts or extracts or translations--are produced. In the case of fact retrieval or question-answering systems, the canonic sentences represent the beliefs of the system and serve as the base for generating factual answers to specific queries. Thus, although most existing automated content analysis procedures are at best low level approximations of this ideal, it is clear that analyzing the content of natural language text must be based on semantic and syntactic principles and is hence an obvious object of linguistic endeavor. It is unfortuante that the significance of this fact is not appreciated by the majority of linguists.

1.1.2 A further--and not unrelated--reason for non-participation linguists in automated language processing is a basic lack of knowledge about computers, in the sense of realizing when a computer is a handy tool, and when it isn't so handy. By this I don't mean a lack of knowledge about hexadecimal systems, bits and bytes, or serial and parallel processors, but very simply knowledge of what a computer is good for.

The fact is that for many of the operations characteristically performed in linguistic research, the computer is an invaluable--if not an indispensable--tool. This is a very strong claim for the utility of the computer in linguistics; therefore, the grounds on which it is based are worth examining in some detail.

The operations which the linguist performs in carrying out research on a language or languages are essentially the following: he collects data, organizes and analyzes them, formulates hypotheses and verifies them. There is of course, a great deal of feedback analysis and recycling through all these operations, which are highly interdependent. It is therefore impractical to examine the applicability of computer processing individually to each of the operations listed above. Since the important concept of "organization" applies equally to data and hypotheses, in the following the linguistic operations for which computers can be used will be grouped into these two categories. Where these operations are differently interpreted or valued by linguists of different schools, divergent points of view will be noted.

Data collection and organization, operations on the data base. There are two senses in which data is collected in linguistic analysis. The first sense refers to the initial collection of data for inclusion in the corpus or data base. For a linguist working with a language unknown to him, this generally means eliciting such data from an informant--a native speaker of the language in question. This operation can not proceed in a haphazard manner, but for the traditional descriptive linguist at least, is one subcomponent of a heuristic strategy for discovering the basic elements and relations of a given language. Because it is a heuristic rather than an algorithmic procedure, explication for a computer is a formidable undertaking. Thus far, there has been only one attempt at automating paradigm elicitation for unknown languages (Garvin 1969); since this project also involves analysis and hypothesis formulation, a more detailed discussion will be presented below.

A second sense of data collection is the selection of particular data items from a previously collected data base. When working with a language in which he does not have native or near native competence, operations on the data base--that is to say, organizing, searching, retrieving, and refiling data--assume a dominant role. This is because the linguist cannot rely on himself as a source of data which he can organize and analyze in terms of his own competence (this procedure presents another type of problem, which is discussed in Section 2). The anthropological linguist thus must devote a disproportionate amount of time simply to operating on the data base, and especially to organizing his data. He typically records his data items on small slips of paper, which he then sorts and cross-files according to various criteria. The problem is that he can only cross-file a data item as many ways as he has duplicate slips, an original and three carbons being about the limit of legibility in recording data with a pen or pencil. The four copies allow him, for instance, alphasorted files of English/Language L, Language L/English and two morphological classifications. If the language is a tone language other than Iraqw--an african language in which tone and morphological classes coincide--four files are insufficient even for morphological analysis; for syntactic and semantic analysis, such a limitation constitutes a serious obstacle.

There are two major problems inherent in these traditional data-handling methods, which may provide at least a partial explanation for the well-known inadequacies in the descriptions of the so-called "exotic" languages (Uhlenbeck 1960). In the first place, the operations involved in the creation of these files, retrieval of relevant data from them, and replacement of the data in the files require a great deal of the linguist's time, which might be more profitably spent in analysis and in hypothesis formulation and verification. Secondly, because in a taxonomic approach, classifications contained in the files in effect form the basis of the grammar, and because syntactic and semantic analysis requires a highly sophisticated and extensive organization of the data, these aspects of linguistic research inevitably suffer when data handling is limited to traditional manual techniques.

Now, the clerical operations of sorting and listing data rapidly and variously are just those at which the computer excels. The computer can speedily present a variety of arrangements of large volumes of data, which may expose underlying patterns not identifiable--or identifiable only with difficulty--by means of traditional card filing techniques. The possibilities for automating these types of operations have yet to be fully exploited; however, programs for generating morpheme concordances have been developed by Grimes and by

Kay (1969).[2]

Formulation, organization, and verification of hypotheses: analytic and synthetic operations. Although a computer cannot spontaneously generate hypotheses, it can assist the linguist in recognizing patterns in the data. Moreover, in organizing and verifying hypotheses, the computer may well be an indispensable tool. In order to test a hypothesis, it should be stated as explicitly as possible; use of the computer forces the investigator to be explicit. In computer testing of hypotheses, loose formulations become obvious rather quickly, as the computer performs all and only the operations specified in the program--often to the dismay of the investigator.

In addition to the stringent requirement for explicitness, use of the computer necessitates a logical organization of hypotheses in order to provide for systematic testing and error tracing. Such requirements apply equally to formal grammars and the somewhat more loosely organized descriptive grammars.

Transformational grammars, however, present a particularly convincing case for the necessity of computer testing. It is difficult to envision how the linguistic researcher can possibly keep track of

---

[2]Although some difficulties are inevitable in converting linguistic materials to machine-readable form, the initial investment of time, energy, and funds are well worth the effort. At present, whether a keypunch or an optical character reader is used as a conversion device, linguistic diacritics and special characters must be recoded in terms of the available character set. However, fully automatic conversion by means of an optical character reader is a development which can be expected within the next few years. Some existing models can read a variety of type styles with the combined error rate of the reader and the typist being lower than that of key punched material, and the recognition of handprinted characters with an acceptable error rate is not far off.

the tortuous ramifications of ordered rules within a single component of the grammar--let alone across components--without the aid of an automated grammar tester. Several computer programs for testing grammar rules have in fact been designed. These include a phonological rule tester (Bobrow and Fraser 1968), several versions of syntactic rule testers based on the MITRE grammar (Friedman 1968; Gross 1968; Gross and Walker 1968), the Transformational Grammer Tester (TGT) developed by Londe and Schoene (1968) for the Air Force UCLA English Syntax Project, and a system constructed by IBM to test the grammar of English II (Rosenbaum 1967). Although these programs all operate through a synthesis procedure, the on-line system described in Gross and Walker also has an analytic capability through the MITRE Syntactic Analysis Procedure.

In addition to these largely synthetic test devices, many analytical algorithms exist. These include algorithms for morphological as well as syntactic analysis. The design of certain types of morphological analysis algorithms for particular languages is in fact fairly well understood. Reasonably successful suffix analysis algorithms have been designed for Russian (Ramo Wooldridge 1960) and for English Chapin 1967; Earl 1967).

Numerous computer programs with various theoretical bases have been designed for analyzing syntax. These include various versions of the Cocke algorithm--a bottom-to-top parsing logic which uses a table of binary IC grammar rules to develop simultaneously all possible analyses of an input string (for a discussion of the Cocke logic, see Hays 1966, pp. 75-7; for recent applications of the Cocke algorithm in automated language processing, see Montgomery 1969). A top-to-bottom predictive equivalent of the Cocke algorithm is the Kuno-Oettinger Syntactic Analyzer (Kuno 1965). Both these algorithms, however, suffer from the disadvantage of producing multiple analyses. More effective procedures for syntactic analysis incorporate transformational rules; these include the MITRE Syntactic Analysis Procedure and that described by Martin Kay (1967). Another approach to syntactic analysis is the "fulcrum" method, reported in Garvin (1968), in which the grammar and the parsing logic are both incorporated into the analysis algorithm.

Although the synthetic rule testing systems discussed above are useful only for testing formal grammars, the analytic algorithms might also be used to test traditional descriptive grammars. In the case of descriptive grammars, criteria for ordering and exhaustiveness are somewhat less rigorously specified than in formal grammar; a descriptive grammar is nevertheless in intricate network of complexly interrelated statements in which some inconsistency is probable if not unavoidable. It would appear that the most effective means of precluding such a possibility is through systematic computer testing.

At present, the most versatile device for testing a descriptive grammar is probably some version of the Cocke algorithm, which could be used as a morphological analyzer with a set of morphological rules, and as a syntactic analyzer with a set of syntactic rules. In morphological analysis, the input string would consist of codes representing the morphs occupying the successive position classes which form the particular word. In syntactic analysis, the input string would of course consist of codes constituting the grammatical labels of the words which form the particular sentence.

Finally, it is appropriate to discuss a computer application which is noteworthy not only by virtue of the fact that it is in the descriptive tradition, but also because it constitutes a substantial departure from the above-mentioned algorithms in several important respects (Garvin 1969). First, both the analytic and synthetic computer systems discussed above are mechanisms for testing a grammar of a particular language; hence, they accept test data and hypotheses in the form of grammatical rules as input and produce as output various diagnostics showing how the data were analyzed by the rules. On the other hand, Garvin's program collects unanalyzed data in an ordered manner by means of its elicitation subcomponent, applies theoretical assumptions to the data, and outputs a hypothesis about the morphological structure of the language represented by the data. Second, rather than testing a grammar of a particular language, the immediate objective of the program is to test a theory of linguistic analysis as represented by a discovery procedure and the ultimate objective is to explicate the universal and near universal assumptions (linguistic universals) that are implicit in the operations of the linguistic analyst. Third, the program thus includes all the operations which a descriptive linguist performs, except for hypothesis verification. Fourth, the computer program is constructed on heuristic, rather than algorithmic, principles.

1.2 The linguist's failure to recognize the significance for linguistics of natural language information is paralleled by the failure on the part of many persons engaged in automated language processing to recognize that the problem is essentially a linguistic one. Moreover, the linguist's lack of knowledge about the computer as a versatile language processing tool is complemented by the lack of linguistic knowledge of his computationally-oriented counterpart.

Examples of non-linguistically oriented computer processing of natural language data in the guise of content analysis are too numerous to mention in detail--the classical example is Luhn's "KWIC" concept (Luhn 1959) and its multitudinous misapplications (for an exhaustive listing of these through 1964, see Stevens 1965). Examples of somewhat more sophisticated approaches include the various attempts to identify concepts and the relations obtaining between concepts without recourse to a systematic syntactic analysis of the given text. Characteristically, the text is segmented into "chunks" or "fragments" by an ad hoc recognition procedure based on lists of prepositions, conjunctions, introductory adverbs, and the like (e.g., Kochen 1969, Bohnert 1966, Briner 1968, Wilks 1968).

1.3 From all the foregoing, the conclusion is inescapable that essentially, the majority of linguists do not have a proper perspective on automated language processing and the majority of non-linguists engaged in automated language processing do not have a proper perspective on linguistics. Nevertheless, these two groups of persons have a common interest in natural language. It is therefore appropriate to examine their perspectives on the nature of language. Generally speaking, their viewpoints tend to be dichotomous; these oppositions underlie the problems discussed above.

2.0 In essence, linguists (especially those of the formal descriptive school) are theory-oriented; persons engaged in automated language processing are data-oriented. Moreover, most linguists would agree that a linguistic theory can be disproved by a single counter-example, no matter how unlikely; whereas researchers in automated language processing are not disturbed by an incompatible piece of data unless its probability of occurrence threatens the practical objective of the application. Linguists search the infinite set of a natural language for their counter-examples while persons engaged in automated language processing pare natural language down to an often skeletal subset, just to exclude data which will perturb their system. Linguists are concerned with the ideal of competence; automated language processing researchers must deal with the facts of performance--"the adulterations of the ideal" (Katz 1967).

If one takes a negative point of view, these dichotomies represent irreconcilable differences in the basic conception of language; more positively, they may be regarded as complementary perspectives on the nature of language. The initial issue is thus one of determining which view is correct. Should the positive view be adopted, there is a more fundamental question as to the potential for unifying the two approaches to provide a balanced attack on problems of natural language analysis and description.

In answer to the first question, it is reasonable to consider the two approaches as complementary, since the specific weaknesses of the data-oriented position are offset by corresponding strengths in the theoretical orientation, and conversely. In the following discussion, the respective deficiencies of the two approaches will be examined and potential unifying concepts will be explored.

2.1 The data-oriented view of natural language is generally characterized by a bias toward the data, a reliance on statistics, an interest in subsets of natural language, and thus a concern with some particular inventory of sentences of performance exclusive of any notion of the infinite inventory of sentences of competence.

There are two general directions in which this weakness is exhibited, depending on the size of the natural language subset that is involved. With extremely large subsets,[3] data orientation is mainly due to data inundation, and computer processing substitutes for theory. In a system of this type, it is possible to perform a great deal of computer processing without knowing quite what it all means. Content analysis may be attempted by statistical techniques, but if the definition of the statistical word is not correlated with an actual word stem--or more relevantly, with a concept which may be represented in natural language text by a number of different words and phrases--then all that has really been performed is a frequency count of unique character strings. The actual process of content analysis remains to be performed.

Another variety of data-orientation weakness involves extremely small subsets of natural language. In this case, the defect consists in the testing of theories on very limited amounts of data--often only

---

[3] In this context, a large subset, is defined as the entire information store in a particular system for processing--say, scientific materials--where the data base consists of over 100,000 documents.

on the very sample from which the theory was originally derived. Claims for the generality of techniques derived by such means must thus be viewed with a certain amount of skepticism. Unfortunately, many of the more interesting activities in automated language processing--e. g. question-answering systems--suffer from this defect.[4]

2.2 On the other hand, there are the weaknesses of the opposite perspective, which is characterized by a preoccupation with theory, counter-examples, and the infinite set of sentences of a speaker's competence. The deficiencies of this approach become apparent in considering a few passages from Katz, excerpted from a polemic between Katz and the philosophers Quine and Wilson.

Referring to a paper in which Quine criticizes Carnap's treatment of analyticity, Katz supports Quine's criticism of Carnap, stating that the Katz-Fodor theory does not require "such ad hoc devices as meaning postulates and semantic rules" to characterize an analytic sentence but rather defines it as "a sentence whose semantically interpreted underlying phrase marker (generated by the optimal grammar for the language) is such that every semantic marking in

---

[4] It is interesting to note in passing that a similar cirticism has frequently been leveled at traditional linguistic descriptions by linguists espousing the generative approach. According to this criticism, the descriptive linguist suffers from an exaggerated dependence on his "corpus"--the body of linguistic material constituting his data base. His description of the language--in formal terms, his theory of the language--is thus a description of the corpus, and its validity is a function of the adequacy of the corpus as a representative sample of the language.

the reading for its predicate also occurs in the reading for its subject" (Katz, 1967). Katz thus defines "S is analytic for L" in terms of theoretical constructs for which he claims universality; he further states that for each language, "for each L_1 that is a possible value of "L", it is possible to differentiate the analytic from the nonanalytic sentences in L_1 on the basis of predictions that follow from this definition in conjunction with the semantic descriptions of the sentences in L_1 provided by the grammar of L_1" (1968).

Unfortunately, the impact of Katz's arguments is substantially reduced by the fact that--although there exists a definition of analyticity which has been postulated by Katz in terms of the theoretical constructs "underlying phrase marker," "semantic interpretation," "reading," "subject of," etc., --there exists no grammar for any L_1 to provide the semantic descriptions of L_1 which must be conjoined with Katz's definition to provide for the differentiation of analytic from non-analytic sentences.

Moreover, if Katz were to state that he had actually produced a grammar of some L_1 complete with semantic descriptions and presumably capable of generating the set of sentences of a speaker's competence in L_1, no one could prove that this was or was not an empty claim. Katz himself has affirmed the necessity of behavioral tests as a means of validating the empirical adequacy of his theoretical formulations (Katz 1967, 1968). However, previous attempts to investigate various syntactic phenomena through behavioral experiments have not been spectacularly successful, and since the investigation of semantic phenomena is inestimably more complex, behavioral verification of a grammar of L_1 appears impossible.

This difficulty derives from two sources, one of which involves the nature of meaning, and the other, the present state of knowledge about linguistic performance, or speech behavior.

The semantic problem lies in the fact that a great deal of meaning is situationally derived; the physical and sociocultural situation to a considerable extent controls the semantic interpretation of sentences. In the narrow sense, the concept of a physical and sociocultural context can be limited to those situations which are participated in by a majority of the speakers of the language: say, a school, a city, an airport. In the broader sense, however, physical and sociocultural context includes such factors as the entire history of an interaction between two persons--in other words, all the occasions on which they have interacted and the content of those instances of interaction. Without such information, a proper interpretation of innuendoes, jokes, allusions, and so forth, would not be possible. Also, in the sense of an interaction between persons, the context is dynamic; it grows from the inception of the interaction to its conclusion.

Thus, the speech event is actually performed in an environment consisting of the entire range of physical and sociocultural phenomena which are relevant to its interpretation. For this reason, semantic interpretation presents problems of considerable magnitude, some of which may be inherently insoluble.

Setting the semantic problem aside for the moment, we consider the second source of difficulty encountered in attempting to validate a grammar of L_1 through behavioral testing: the present lack of an adequate theory of performance, or speech behavior. A grammar is a model of a speaker's innate capacity, and not of the ways he uses this capacity to produce and understand sentences. Although experiments suggest the psychological reality of some features of the structural descriptions generated by the competence model or grammar (Fodor and Garrett 1967), a speaker demonstrates his competence through his performance, and the relation between a speaker's competence and his performance has yet to be explicated. Assuming that a speaker of L_1 will produce and understand only sentences for which the grammar of L_1 can supply structural descriptions, the problem is reduced to determining how the speaker behaves in terms of the structural description, which is not trivial to begin with.

However, reintroducing the semantic problem discussed above, it is clear that the explication of performance involves specification of the speaker's behavior in composing and interpreting sentences with respect not only to structural descriptions, but also to the total environment of the speech event. Thus speakers can and do process sentences which the grammar is not capable of generating; in other words, the relation between the sentences of competence and those of performance is not one of simple inclusion. As noted by Kasher (1967) and developed in detail by Watt (1968), there are certain features of the sentences of performance which cannot be replicated in a competence model--these include those which are derived in some way from the physical and sociocultural situation. An example of such a feature is deletability; the sentences of performance are characterized by deletions which are not recoverable from the immediate linguistic context, but must be supplied from the physical and sociocultural environment. Unfortunately, formal grammars tend to be based on isolated examples of the performance of the linguistic investigator, rather than on spontaneous speech. This practice has the disadvantage of effectively eliminating examples of speech which depend for their interpretation on the total environment of the speech event. For instance, the cryptic statement "Number Five once" is not mysterious at a race track, where the numerous deletions are recovered from the environment (the $2 Win window of a thoroughbred race track) to provide something like the following:

(a) "I would like to wager two dollars of the five dollars in my hand that the horse which is starting at Post Position Five will win the next race."

Because the sentences of performance are largely context-dependent, and because there is as yet no explication of how speakers behave in terms of structural descriptions--let alone in terms of the total environment of the speech event, it is apparent that the majority of sentences produced by speakers of L_1 could not be generated by the grammar of L_1. Thus, a grammar of L_1--assuming the existence of such--could not be validated by behavioral tests, and there would be essentially no way of relating the sentences actually performed by speakers of L_1 to those specified by the grammar. It is therefore appropriate to inquire what such a grammar might be good for.

From the data-oriented point of view it is clearly inadequate, because it does not deal with the sentences actually produced by speakers of the language; from the theory-oriented point of view it is also unsatisfactory, since it is incomplete. Yet, because of the improbability of explicating the total environment (as defined above) of the speech event, there is little hope that a complete semantic theory will ever be developed, and full understanding of how a speaker uses the competence which the grammar represents is not likely to be acquired in the foreseeable future. It therefore seems worthwhile to consider some concepts and strategies which might serve as working hypotheses and provide at least an interim solution to problems of the theories of meaning and speech behavior.

2.3 One concept which might prove useful in this regard is that of "semantic equivalence". Returning to the context-dependent example of race track parlance presented above, it is clear that given the physical environment of the track, past experience in that environment, and other relevant sociocultural phenomena, the speaker of American English accepts "Number Five once" as in some way equivalent to the explicit proposition presented in (a). The two examples belong to a set of sentences which might be described as "semantically equivalent performances" for the purposes of this presentation. These sentences are thus defined on the assumption that speakers of American English would judge them to function as semantic equivalents in the appropriate environment (provided that the speakers were knowledgeable about the particular environment, either vicariously or through personal experience. Some additional members of the sample set are as follows:

"I want to bet two dollars on Number Five."

"I want Number Five once."

"Give me a two dollar bet on the Number Five horse."

"Put two bucks on Five."

"Two on Number Five."

Note that some of these sentences would be judged syntactically deviant by speakers whose experience does not include participation in the milieu of a race track. Moreover, the majority of the sentences would be judged semantically deviant in a non-betting environment. Those sentences which would be judged semantically appropriate in other environments, however, derive their appropriateness not from membership in the set presented above, but from membership in some other equivalence set which is semantically appropriate for the given environment. For example, "Two on Number Five" might also be used in an airport; but in this case, it would belong to an equivalence set including the following sentences, among others:

"I want two seats on Flight Number Five, which leaves Great Falls at 6:05 a.m. and arrives in Salt Lake City at 8:49 a.m."

"Give me two tickets on the flight that leaves here at 6:05 a.m."

"I want two seats on Number Five to Salt Lake City."

"Two tickets on the next flight to Salt Lake City, please."

In considering the notion of equivalence set as a possible working hypothesis, a few operating difficulties should be noted. One such difficulty lies in explicating the definiens or basic member of the set, since it must include all the relevant features of the physical and sociocultural environment. A second problem consists in defining appropriate equivalence sets for more abstract contexts, where the notion of equivalence is more difficult to specify than in the examples presented above.

However, this notion--which defines sets of instances of behavior in terms of their function as semantic equivalents in particular physical and sociocultural environments--is useful for two reasons. In the first place, it provides a means for dealing systematically with the elusive concepts of speech behavior and situationally derived meaning. Secondly, the notion of an equivalence set provides an approximate definition of a relation, in the sense of symbolic logic, and is thus a means of approaching a formalism in an inductive way.

The difficulties which inhere in the explication of meaning and of speech behavior make it rather unlikely that such theories will spring fully developed from the brow of some linguistician. Therefore, if complete explication of meaning and speech behavior is possible at all, it would seem more likely to be achieved by working from the explicit to the inexplicit than conversely.

Accordingly, it is suggested that a reasonable approach to problems of the theories of meaning and of speech behavior would be the construction of an experimental model for analysis of natural language in terms of sets of semantically equivalent performances as defined above. The initial model would be developed from a data base consisting of sentences performed in a particular physical and sociocultural environment, and would thus represent a restricted subset of the natural language. The environment selected for the original model might be a race track, an airport, a market, or some other type of structured situation, in order to reduce problems of defining semantically equivalent sets of sentences. Successive versions of the model would be capable of processing materials of increasing complexity with respect to contextual variables--e. g. the various subsets of "present-day American English" represented in Kučera and Francis (1967).

Assuming a restricted automatic thesaurus and a data base in machine-readable form, a first cut at equivalence sets could be provided by separate lists(sorted internally by number of thesaurus group assignments) of sentences containing words or phrases from the same thesaurus groups, and words and phrases from the same group as well as more general or more specific groups. These lists could then be studied in detail to isolate potential equivalence sets. The elements of the basic member or definiens of each set would be identified in the course of this study, and the set membership validated by behavioral tests, which would also serve as a means of eliciting additional members of the set not represented in the data base.

The final step in construction of the model consists in representing the definiens in the notation of formal logic, and representing the other members of the set in terms of the definiens. Analysis of a sentence presented to the model is thus accomplished through a decision procedure for membership in a particular equivalence set, by association with a particular definiens or its converse.

3.0 The proposed model is presented as an approximate solution to problems of theory and data orientation. It overcomes the respective weaknesses of the two approaches (see Sections 2.1 and 2.2) by providing a means of arriving at theories of meaning and speech behavior through exploitation of data bases which are subsets of a natural language containing instances of speech behavior used in particular physical and sociocultural environments. Moreover, the concept of equivalence set provides a data defined approximation of the theoretical notion of a relation, in the sense of symbolic logic. This is of particular interest because symbolic logic has been used as a system of semantic representation both in computer processing of natural language data (Montgomery 1969, especially question-answering systems) and in linguistics (McCawley 1969). Some convergence of linguistic and computational viewpoints is thus already in evidence. If progress toward the explication of natural language and the operations involved in processing it (whether by men or machines) is to continue, linguistic science and automated language processing must increasingly share theories and data, objectives and methods.

BIBLIOGRAPHY

BOBROW, DANIEL G.; FRASER, J. BRUCE. A phonological rule tester. Communications of the ACM, 11:11 (November 1968) 766-772.

BOHNERT, HERBERT G.; BACKER, PAUL O. Automatic English-to-logic translation in a simplified model. A study in the logic of grammar. Final report, 1961-1966. IBM Watson Research Center, Yorktown Heights, N. Y., March 1966, 117 p. (AD-637 227).

BRINER, L. L.: CARNEY, G. J. SYNTRAN/360, a natural language processing system for preparing text references and retrieving text information. IBM Corp., Gaithersburg, Md., 1968. (Preprint)

CHAPIN, PAUL G. On the syntax of word-derivation in English. MITRE Corp., Bedford, Mass., September 1967, 191 p. (Information system language studies, no. 16) (MTP-68) (MITRE Project 1117)

EARL, LOIS L. Automatic determination of parts of speech of English words. Mechanical Translation and Computational Linguistics, Vol. 10, nos. 3 and 4, September and December, 1967. pp. 53-67.

FODOR, J. A.; GARRETT, M. Some reflections on competence and performance. In: Lyons, J.; Wales, R. J., eds. Psycholinguistics papers. Chicago. 1967, p. 135-154

FRIEDMAN, JOYCE. A computer system for writing and testing transformational grammars. Final report. Standord University, Department of Computer Science. Stanford, Calif., September 1968. 14 p. (CS-109)

GARVIN, PAUL L. Simulation and analysis of intelligent behavior. Preprint for Wenner-Gren Symposium on Cognitive Studies and Artificial Intelligence Research, University of Chicago, March 2-8, 1969. 23 p.

GARVIN, PAUL L. The place of heuristics in the fulcrum approach to machine translation. Lingua. 21 (1968) 162-182.

GROSS, LOUIS N. A computer program for testing grammars on-line. MITRE Corp., Bedford, Mass., July 1968, 63 p.

GROSS, LOUIS N.; WALKER, DONALD E. On-line computer aids for research in linguistics. To appear in Proceedings of the IFIP Congress, Edinburgh, 1968. North Holland Publishing Co., Amsterdam. In press.

HAYS, DAVID G. Readings in automatic language processing. American Elsevier, New York, 1966. 202 p.

KASHER, ASA. Data-retrieval by computer: a critical survey. In: Kochen, Manfred, ed. The growth of knowledge: Readings on organization and retrieval of information. Wiley, New York, 1967. pp. 292-324.

KATZ, JERROLD J. Some remarks on Quine on analyticity. Journal of Philosophy, 64 (February 1967) 35-52.

KATZ, JERROLD J. Unpalatable recipes for buttering parsnips. Journal of Philosophy, 65:2 (January 1968) 29-44.

KAY, MARTIN. The computer system to aid the linguistic field worker. Presented at the annual symposium of the Interamerican Program on Linguistics and Language Teaching, Sao Paolo, Brazil, January 9-14, 1969.

KAY, MARTIN. Experiments with a powerful parser. RAND Memorandum RM-5452-PR, the RAND Corporation, Santa Monica, California, October 1967. 28 p.

KOCHEN, MANFRED. Automatic question-answering of English-like questions about simple diagrams. Journal of the Association for Computing Machinery, 16:1 (January 1969) 26-48 (AD-670 545)

KUČERA, HENRY; FRANCIS, W. NELSON. Computational analysis of present-day American English. Brown University Press, Providence, R.I., 1967, 424 p.

KUNO, SUSUMO. The predictive analyzer and a path elimination technique. In David G. Hays, Readings in automatic language processing. American Elsevier, New York, 1966. pp. 83-106.

LONDE, DAVE L.; SCHOENE, WILLIAM J. TGT: Transformational grammar tester. In: AFIPS conference proceedings, vol. 32, 1968 Spring Joint Computer Conference. Thompson, Washington, D.C., p. 385-393.

LUHN, H. P. Keyword-In-Context index for technical literature (KWIC Index). Report no. RC 127, International Business Machines Corporation, Yorktown Heights, New York, 1959. 16 p.

MC CAWLEY, JAMES D. Semantic representation. Preprint for Wenner-Gren Symposium on Cognitive Studies and Artificial Intelligence Research, University of Chicago, March 2-8, 1969. 30 p.

MONTGOMERY, CHRISTINE A. Automated language processing. In Annual Review of Information Science and Technology, vol. 4, Carlos A. Cuadra, ed. Encyclopedia Britannica, Inc., Chicago. (In press).

RAMO-WOOLDRIDGE, a Division of Thompson Ramo Wooldridge, Inc. Machine translation studies of semantic techniques. (AF30(602)-2036) Technical Report No. 1 to Rome Air Development Center, Griffiss AFB, New York. Los Angeles, California, 22 February 1960. 142 p.

ROSENBAUM, PETER S. Specification and utilization of a transformational grammar. Scientific report no. 2, October 1966-September 1967. IBM Watson Research Center, Yorktown Heights, N. Y., October 1967, 272 p. (AFCRL-68-0070) (AD-667 800)

STEVENS, MARY ELIZABETH. Automatic indexing: a state-of-the-art report. NBS Monograph 91, National Bureau of Standards, U. S. Department of Commerce, March 30, 1965. 220 p.

UHLENBECK, E. M. The study of the so-called exotic languages and general linguistics. Lingua 9, 1960. pp. 417-34.

WATT, W. C. Habitability. American Documentation, 19:3 (July 1968) 338-351.

WILKS, YORICK. Computable semantic derivations. Systems Development Corp., Santa Monica, Calif. 15 January 1968, 160 p. (SP-3017)