Cathode-Ray Tube

The letters we have generated were drawn on a specific machine, the Stromberg-Carlson 4020 microfilm printer. A brief description of the SC 4020 will assist in using the vector letters with other machines. The SC 4020 has two modes of operation. First, it can draw vectors, which can start at any raster point on its 1024 × 1024 grid and extend up to 64 grid spaces in either or both x and y directions. Secondly, it can produce a total of 64 different characters by shaping the electron beam with an appropriate mask in the cathode-ray tube. One character is a dot. This mode allows one to construct shapes using closely spaced dots, or any other available character, as building blocks. In the type fonts described here, only the vector mode of operation was used. Measurement of the width of the vector indicates that it is equal to 2.3 grid spaces. This means that a character which is 23 grid spaces high has a resolution of only 10 vector widths.

Conclusions

The three fonts of letters presented here are the beginning of a great variety of possible fonts and characters which will be numerically described and computer drawn. The generality of the representation is clear from the ease with which the vectors can be adapted to other computers and other cathode-ray tubes. We believe the fonts will have great utility.

A Grammar Base Question-Answering Procedure

PETER S. ROSENBAUM
IBM Watson Research Center
Yorktown Heights, N. Y.

The subject of this paper is a procedure for the automatic retrieval of certain segments of stored information, either explicitly or implicitly represented, through questions posed in natural language sentences. This procedure makes use of a sentence recognition device for the class of grammars which will correctly decide between the grammatical and ungrammatical sentences of a natural language. It is possible to make use of a recognition device of this sort for the following reason: Much data is fully expressible as a set of sentences in a natural language, a set which can be exhaustively and exclusively generated by a grammar. Based upon the rules of this grammar, a sentence recognizer will evaluate sentences, questions in the normal situation. Since the recognition function succeeds just in case the posed question is drawn from the set of sentences expressing the data, or, more correctly, is grammatical in terms of the grammar for this set of sentences, sentence recognition itself is a procedure for retrieving information. When the recognition function succeeds, its value represents the requested information.

This work was partially supported by the Air Force Cambridge Research Laboratories under contract AF19(628)-5127

1. Introduction

> Does the train which leaves from Croton at 4:10 for New York stop at Yonkers?

According to the Hudson Division Timetable of the New York Central Railroad, the answer to this question is "yes." Establishing this timetable as the definitive data base, we may find it of interest to determine whether the answer to this question can be computed directly or indirectly from the input question itself. An initial assumption is made here that the syntactic analysis component of such a question-answering procedure (a component which is generally agreed to be a necessary ingredient of question-answering systems of more than marginal sophistication) is transformational in nature. This assumption requires that sentences are assigned semantically interpretable deep structures which are mapped onto surface structures through transformational processes. Since, of the grammatical descriptions of English proposed and developed to date, only transformational grammars display even a partial capability of providing synonymous variants of sentences with canonical representations (i.e., common deep structures), the adoption of a transformational syntactic component seems justified. To illustrate synonymous variants, one might cite the sentences below, sentences which are synonymous with the original sentence above but by no means exhausting the set of possible synonymous variants.

> Does the 4:10 train which leaves Croton for New York stop at Yonkers?
> Is it the case that, leaving from Croton at 4:10, there is a train for New York which stops at Yonkers?
> Does the 4:10 from Croton to New York make a stop at Yonkers?

Most commonly, question-answering procedures have been contemplated in which the grammar and the data

base are independent and related by a procedure which translates the output of syntactic analysis into operations to be performed on the data base. Such systems could be referred to as *data base question-answering procedures.* One can imagine a somewhat different sort of question-answering procedure in which, based upon deep structures assigned to input sentences by a recognition routine accepting a descriptively adequate class of grammars, the data base itself is the grammar. The purpose of the following sections is to sketch the salient properties of such a procedure, one which could be thought of as a *grammar base question-answering procedure.*

2. Deep Structure

The deep structure assigned to the interrogative sentence

> Does the train which leaves from Croton at 4:10 for New York stop at Yonkers?

by the most recent version of the IBM English Grammar [1] is very roughly the structure specified in Figure 1.

The syntactic criteria which justify this structure cannot, because of time limitations, be elucidated here. Still, certain general remarks will render this deep structure more informative than it is likely to appear at first glance. Notice, first, that the top level sentence, S_1, consists of: (1) a subject, which is a noun phrase containing the syntactic material necessary to generate a string like "the train which leaves from Croton for New York at 4:10," (2) a predicate, which is a verb phrase interpretable as "stops at Yonkers," and (3) a constituent, Question, which marks the interrogative nature of the top level sentence.

Consider now the structure of the subject of S_1, namely NP_1. This noun phrase consists of a noun phrase head, "train," and a complement sentence interpretable as the relative clause "which leaves from Croton for New York at 4:10." The structure of the relative clause sentence, S_2, is particularly important. This sentence consists of: (1) a subject, which is a complex noun phrase, and (2) a predicate interpretable as "at 4:10." The structure of the subject noun phrase gives rise to the interpretation implicit in the string "that the train leaves from Croton for New York." This treatment allows the explanation of the synonymy of the two sentences below.

> The train leaves from Croton for New York at 4:10.
> It is at 4:10 that the train leaves from Croton for New York.

Common to both sentences is a deep structure in which the subject noun phrase consists of a head noun "it" and a complement sentence "that the train leaves from Croton for New York." The first sentence results from a transformational reduction of this structure. The second results simply from the transformational extraposition of the complement sentence.

Finally, observe that the complement sentence, S_3, originates as the conjunction of two sentences, S_4 and S_5. Aside from intuitive justification, there are many facts

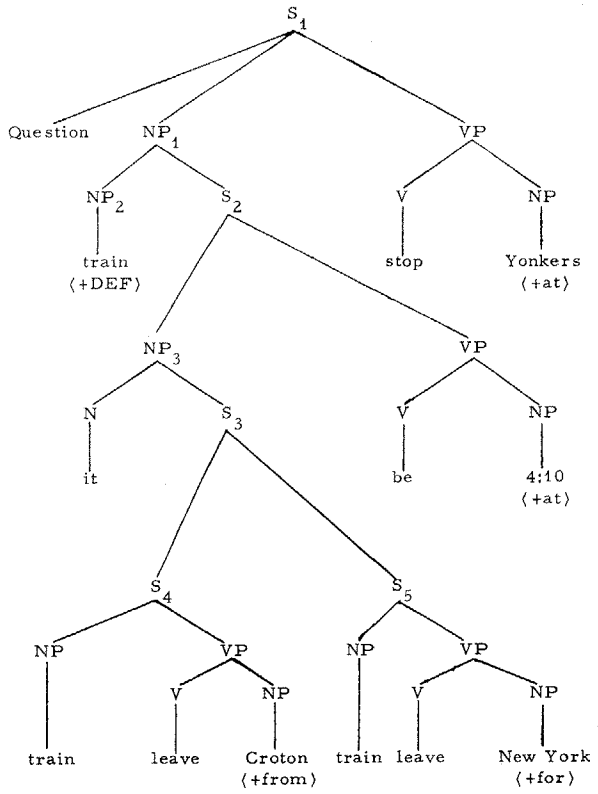


FIG. 1

which lead to the conclusion that a sentence like "the train leaves from Croton for New York" originates as the conjunction of "the train leaves from Croton" and "the train leaves for New York." First, the compound sentence

> The train leaves from Croton and the train leaves for New York.

exists and is synonymous with the reduced sentence below.

> The train leaves from Croton for New York.

Second, the conjunction analysis would explain the free order of "from Croton" and "for New York." Third, such an analysis would allow restrictions obtaining between the verb "leave" and the phrases whose pro-forms are "from somewhere" and "to" or, equivalently, "for somewhere" to be stated simply as verb-object restrictions.

Thus, we arrive at an abstract deep structure which characterizes a number of important syntactic and semantic facts about the original sentence and many of its paraphrases. Although no transformational rules have been discussed, rules which generate surface structures based upon this deep structure nonetheless exist and can be examined in the literature [1, 2].

3. Grammar Base Question-Answering Procedure

The grammar base question-answering procedure does not answer questions at all, strictly speaking. Rather, it determines whether there exists a sentence which consti-

tutes a correct answer to a question. In doing this, the grammar base procedure makes use of one central characteristic of a transformational grammar, namely, its ability to distinguish between the grammatical sequences of a particular natural language and the ungrammatical sequences.

A basic aim of linguistic analysis is to separate the grammatical sequences which are sentences of a language from the ungrammatical sequences which are not sentences of the language. In linguistic analysis, no grammatical sequence of a particular natural language is precluded as evidence confirming or disconfirming, as the case may be, various aspects of a proposed grammar. But it is important to be aware that the notion *grammatical sentence* holds equally well for a subset, either finite or infinite, of the sentences in a natural language. Herein lies the connection between natural language syntactic analysis and natural language question-answering.

Imagine a language composed of the set of sentences A including sentence (1), all English paraphrases of (1), and no other sentences. The set A will contain, consequently, only interrogative sentences of the "yes-no" variety.

> Does the train which leaves from Croton at 4:10 for New York stop at Yonkers? (1)

The notion *grammatical sentence* can be defined in terms of A such that all and only the sentences of A are grammatical. The grammar of A, therefore, will separate the sentences of A from those which are not sentences of A. The latter set will, of course, contain an infinite number of sequences which are grammatical sentences in full English, but this is immaterial. Suppose, now, that the value "yes" is consistently and uniquely associated with all members of A. Under such circumstances, the judgment "grammatical" supplied by the grammar of A for a particular sentence is equivalent to the value, or more appropriately, the answer "yes." It is seen that a grammar which is capable of differentiating the sentences of A from those not of A is also capable, albeit indirectly, of supplying answers to the questions which make up the set A. Thus, answering the questions in A requires no mapping from the output of the grammatical analysis of an input sentence onto a data base since acceptance of the sentence as grammatical implies that the answer to the question is "yes." In short, in the grammar base question-answering procedure the grammar, itself, is the data base.

We point out here that the successful utilization of such a system in any serious application will require that the grammar assign canonical representations to all sets of sentences which are synonymous on a certain reading. Of practical significance, therefore, is the fact that this requirement at present exceeds the capability of any available linguistic description, the notable successes of transformational analysis in this area notwithstanding. Indeed, Chomsky's [3] notion of "deep structure" itself has been challenged [4] on the grounds that it does not provide the basis for a correct description of synonymy relations. These considerations suggest that, with the possible exception of pilot development, the successful implementation of the grammar base question-answering proposal will be hampered by rather severe linguistic limitations and that an effort in computational research must be paralleled by a commensurate effort in linguistic research.

The grammaticality or ungrammaticality of a particular deep structure, with respect to the grammar base question-answering procedure, is largely a function of what might loosely be called its *selectional well-formedness*. This notion is intended as an evaluational term reflecting facts like those which emerge upon consideration of the strings below.

> The professor smiled.
> * the collision smiled

One aspect of a native speaker's linguistic knowledge of English is the knowledge that these two strings have a different status. Consequently, a grammatical description claiming to explicate this knowledge must explicitly differentiate between these two strings and must, furthermore, assign a degree of deviation, or ungrammaticality, to the second string. In the most recent formulation of linguistic theory, [3] nouns, within a simple sentence, are said to exercise selectional power over the main verb of the sentence. Thus, if the subject noun is human, (indicated by the binary feature ⟨+human⟩), the main verb node, V, is subcategorized as requiring a human subject. Furthermore, individual lexical items are specified, in part, in terms of their unique selectional properties. For example, the verb "smile" is lexically represented with the *selectional feature* ⟨+ ⟨+human⟩ ___ ⟩, which asserts that this verb is positively specified for its possibility of occurrence following a human subject. Such a selectional mechanism is sufficiently rich to explain the difference between the two strings singled out above. For the first string, "the professor smiled," a human noun has subcategorized the verb node in the deep structure as requiring a human subject and, moreover, the lexical item "smiled," which possesses exactly this selectional property, was introduced into the deep structure under the domination of the verb node. On the other hand, for the string "*the collision smiled," the verb is subcategorized as taking an abstract subject. Since the subcategorization imposed by the subject noun conflicts with the selectional properties specifying the privileges of occurrence of the lexical item "smiled," this string is, predictably, selectionally ill-formed.[1]

The selectional mechanism is the critical component of the grammar base question-answering procedure. Consider again the grammar of A. This grammar will have an associated lexicon in which the selectional properties of verbs, and of predicates generally, with respect to the grammatical sentences of A will be fully specified. The lexicon for A will contain basically two entries for the verb "leave." The first entry will stipulate that a possible object

[1] This presentation of the selectional mechanism is necessarily much oversimplified. For a more detailed discussion see [3].

noun of this verb is one specified as ⟨+Croton⟩ ⟨+from⟩; the second entry will stipulate a possible object as ⟨+New York⟩ ⟨+for⟩. The subject for both verbs would be specified as ⟨+train⟩. Any interrogative sentence, such as

Do trains leave from Croton?

which reduces to a deep structure in which the selectional properties of the verb agree with the subcategorization imposed by the subject and object nouns will be selectionally well-formed, the consequence of which is the answer "yes." Sentences, such as

Do trains leave from Armonk?
Do buses leave from Croton?

will reduce to a deep structure in which the properties of "leave" conflict with the subcategorization imposed by the subject and object nouns. Such deep structures, from the point of view of the grammar of A, will be selectionally ill-formed, the consequence of which is some appropriate answer other than "yes," (e.g., "no," "I don't know," etc.).

Selectional well-formedness, in the sense of this discussion, must be determined recursively, a fact which has bearing both on the operation of the system and on the structure of the lexicon. For illustrative purposes, consider the sentence

Does a train from Croton stop at Yonkers?

which has roughly the deep structure given in Figure 2.

In earlier examples, the subcategorization of predicates was specified fully in terms of the simple sentence in which the predicates occurred. For example, for the sentence

The boy smiled.

the subcategorization is entirely a function of the inherent properties of the subject noun, "boy." However, this does not appear to be the case in Figure 2, since the subject noun "train" is not sufficiently specified to subcategorize the predicate "stop" correctly. The lexical entry for "stop" must specify as a possible subject just that train which leaves from Croton.

The content of embedded sentences is thus seen to be relevant to the subcategorization of verb nodes in higher sentences. Even given a highly circumscribed data base, this circumstance creates enormous lexical complications since it is no longer possible to state the privileges of occurrence of verbs simply in terms of the nouns which surround them locally. Now it appears necessary to state, in the lexicon, the occurrence of verbs with respect to sentences embedded in the subject and object noun phrases and, even worse, with respect to sentences embedded in these sentences and so on and on and on. The inclusion of syntactic structure in lexical entries is not only a major complication from the practical point of view; it fails to capture an important generalization, namely, that it is the semantic content, if you will, of such embedded sentences and not the syntactic content which is relevant to stating privileges of occurrence. Nonetheless, we seem to

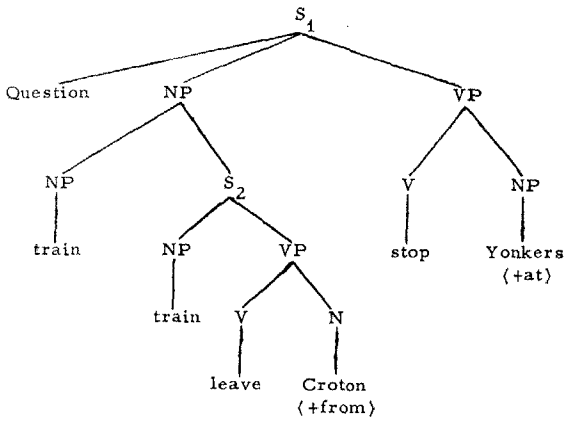


FIG. 2

be forced to the regrettable conclusion that syntactic structure, for embedded sentences, must be included in lexical entries. (It is scarcely necessary to mention here the fact that such a degeneralization of the grammar will have commensurate effects upon the operation of the recognition procedure which, in determining selectional well-formedness, will have to process trees rather than strings.)

Several observations lead to an interesting resolution of the problem of recursion. First, as previously mentioned, the subcategorization of the predicate "stop" is in part determined by the relative clause embedded in the subject noun phrase. Second, consider the fact that the relative clause itself is fully represented by the lexical entry for "leave." This follows from the fact that the selectional well-formedness of the relative clause implies the existence of a lexical entry for "leave" which allows "train" as a subject and "from Croton" as an object. Thus, the entire relative clause is uniquely represented by the lexical entry for the verb "leave" in that relative clause. In a very real sense, it appears that the entire sentences are representable as words (i.e., lexical entries). Since all lexical entries will be distinct (to the extent that they are not synonymous), the content of a lexical entry can be abbreviated by some inherent feature, let us say an integer. The following lexical entries are illustrative.

leave: [⟨+V⟩ ⟨+ ⟨+train⟩____⟨+Croton⟩⟩ ⟨+from⟩ ⟨+32⟩]

leave: [⟨+V⟩ ⟨+ ⟨+train⟩____⟨+New York⟩⟩ ⟨+for⟩ ⟨+41⟩]

The effect of representing embedded sentences as distinct lexical entries becomes clear when it is observed that the privileges of occurrence for the verb "stop" in Figure 2 can now be stated without including the constituent structure of the relative clause in the lexical entry itself. The

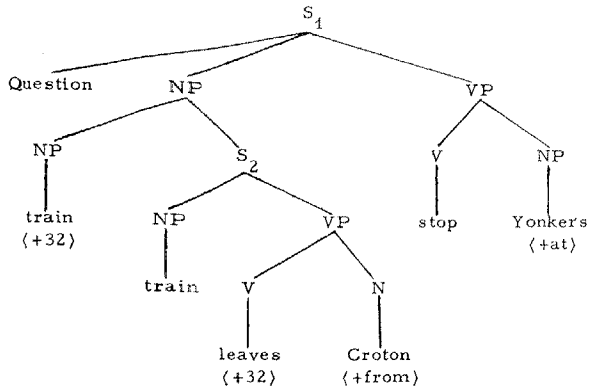


FIG. 3

subject of "stop" must simply be a noun with the feature ⟨+32⟩, as is represented in the following lexical entry.

$$\text{stop}: \begin{bmatrix} \langle +V \rangle \\ \langle +\ \langle +32 \rangle \_\_ \langle +\text{Yonkers} \rangle \rangle \\ \qquad\qquad \langle +\text{at} \rangle \\ \langle +96 \rangle \end{bmatrix}$$

The only question remaining concerns how the feature ⟨+32⟩ on the verb in the embedded sentence is generated on the head noun "train" appearing in the top level sentence as the subject of "stop."

Imagine a procedure for determining the selectional well-formedness of deep structures in a cyclic fashion, that is, a procedure which analyzes deep structures sequentially beginning with the most embedded sentence (or sentences) and proceeding higher until the highest sentence, the non-embedded sentence, is processed. Suppose, furthermore, that this procedure consists of two operations. The first compares the subcategorization requirements with the selectional properties of the predicate. If the sentence is selectionally well-formed, the second operation assigns the defining feature of the predicate, e.g., ⟨+32⟩ in the processing of the relative clause of Figure 2, to the head noun of the noun phrase in which the embedded sentence appears. Thus, the first cycle of operations to the structure in Figure 2 generates the structure given in Figure 3. The second cycle of operations now tests the selectional well-formedness of the highest sentence, S_1. Since there exists a lexical entry for "stop" which allows a subject noun with the feature ⟨+32⟩ and an object noun with the features ⟨+Yonkers⟩ ⟨+at⟩, the sentence will be evaluated as selectionally well-formed.

Customarily, interrogative sentences are classified either as *yes-no questions* or a *WH questions*, a distinction illustrated by the two sentences below.

Does a train from Croton leave for New York at 4:10?
When do trains leave from Croton for New York?

WH questions, as is seen, constitute a request for the selectional properties of particular predicates. Recall, for example, that in the language given by the subset of sentences A, the verb "leave" has, as one of its possible objects, a noun with the features ⟨+Croton⟩ ⟨+from⟩. Thus, the query

Where do trains leave from?

amounts to a request for the specification of the possible objects of "leave" marked ⟨+from⟩, which, of course, are given in the lexical entry for "leave." For A, this query would receive the answer "Croton" and only "Croton."

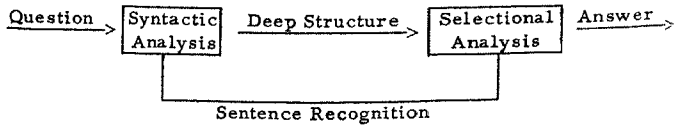
Basically, the development of a grammar base question-answering system requires the construction of: (1) a grammar for the appropriate data base and (2) a recognition routine. The grammar must contain a transformational component and a lexicon. Inasmuch as there are no known discovery procedures for tranformational grammars, the grammar must be developed "by hand," as it were, although computational aids in the form of sentence synthesizers which test the generative adequacy of the grammar will certainly facilitate the process.[2] The lexicon, on the other hand, can, in large measure, be generated automatically on the assumption that the recognition device will interpret a declarative sentence as a lexical update instruction rather than as a query. For example, take the sentence

Trains leave from Croton.

The recognition device generates a deep structure. This deep structure, similar to the embedded sentence in Figure 2, determines a subcategorization for the verb "leave." A lexical entry is automatically constructed in accordance with this subcategorization. If the new entry is distinct from all other entries in the lexicon, it is assigned an integer feature, e.g., ⟨+32⟩, and entered in the lexicon. If it is not distinct, the entry is synonymous with an already present entry and, consequently, ignored. Correspondingly, entries are deleted from the lexicon simply by supplying the system with the appropriate declarative negative sentence, e.g., "Trains do not leave from Croton." The recognition routine itself for the grammar base question-answering system (1) must be capable of generating deep structures for grammatical surface structures and (2) must be capable of evaluating selectional well-formedness in the cyclic fashion suggested above.

4. Conclusion

The grammar base question-answering procedure proposed in the preceding discussion fits into the following paradigm.



[2] Such a synthesizer was developed at IBM to evaluate the core grammar referred to in [2] and is described in [5].

Since the system under discussion is not yet an operating system, it is somewhat difficult to determine its potentialities. Still, certain of its more general characteristics can be enumerated.

1. The system will accept complex (i.e., embeddings) and compound (i.e., conjoined) input sentences.
2. The system will accept sentences with many-place predicates (which are reduced to one or at most two-place predicates in deep structures).
3. The lexicon can be developed and altered automatically through submission of English sentences to the system with minimal regard to format.
4. It should be possible to update and indefinitely extend the grammar without any reprogramming of the recognition routine.
5. The number of relations, i.e., predicates, which the system will accept is probably unlimited. The class of relations is undoubtedly restricted, but such restrictions stem more, in all likelihood, from the paucity of confirmed results in linguistic analysis than from a deficiency in the system.
6. The system will require a lexicon which is extremely large relative to the size of the formatted data base upon which the system is based.

The grammar base question-answering system is not proposed as a cure-all for any or all of the practical problems which confront those researchers currently involved in the development of question-answering systems for production purposes. Such systems, if sufficiently limited, can no doubt make do with a linguistic analysis routine which is considerably less sophisticated than the one proposed here and, for various reasons, the use of less powerful routines is probably well advised. However, discussions of artificial limitation of a natural language must proceed largely in a vacuum at the present since virtually nothing is known either about the intuitive linguistic preferences of a speaker confronted with a question-asking assignment with respect to a given data base or about the learnability of subsets purporting to be based upon a grammar describing synonymy relations within such subsets. On the latter, see [6].

Nor is the proposed system being promoted as the only possible system for answering questions on the basis of a transformational analysis of sentences. One can easily imagine a more conventional data base question-answering system in which the data base is kept distinct from the grammar and which requires procedures for mapping deep structures onto this data base for the purpose of getting an answer. Rather, the proposal under discussion should be viewed as a promising unique approach for the general problem of question-answering possessing unusual design features which are directly a consequence of linguistic descriptions developed in accordance with the transformational theory of syntax. As such, it is should be of general interest to explore the characteristics of the system and its potentialities further, with respect both to pilot applications and to advanced linguistic research in abstract syntax. Such an attempt will bring to light a number of fundamental questions which are neither answered or even raised in the present paper. Illustrative of these are the following: First, how are negative bits of information to be handled? For example, if a train stops at Yonkers every day except Sunday, how will this information be incorporated into the system? Second, how will attributive relationships be treated? In the sentence," Iron is a metal," should it be assumed that "be" is a verb with selectional restrictions pairing "iron" with "metal" as subject and object or will "iron" be entered in the lexicon with a property list and "be" be interpreted as an instruction to scan the property list? Third, how are relevant questions to be distinguished from irrelevant questions? In other words, what determines when the system answers "no" and when it answers, in effect, "I don't know"? Questions such as these are current research topics and the answers to them, at the time of writing, are too sketchy and inconclusive to elaborate. Suffice it for the present that studies of the problems of question-answering inherent to the grammar base question-answering procedure are currently being undertaken. A detailed evaluation of the system's merits and liabilities can be expected in the future.

Acknowledgments. The author is grateful to Dr. Warren Plath and Dr. Jane Robinson who read earlier versions of this manuscript and provided numerous helpful suggestions.

RECEIVED JUNE, 1967

REFERENCES

1. ROSENBAUM, P. S. English grammar II. IBM Corp., Yorktown Heights, N. Y. (In preparation)
2. ROSENBAUM, P. S., AND LOCHAK, D. The IBM core grammar of English. In "Specification and utilization of a transformational grammar," Scientific Report No. 1, IBM Corp., Yorktown Heights, N. Y., 1966.
3. CHOMSKY, N. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, Mass., 1965.
4. LAKOFF, G. Instrumental adverbs and the concept of deep structure. Department of Linguistics, and the Computation Lab., Harvard University, Cambridge, Mass., Mimeograph, 1967.
5. ROSENBAUM, P. S., AND BLAIR, F. In "Specification and utilization of a transformational grammar," Final report, IBM Corp., Yorktown Heights, N.Y., 1966.
6. ROSENBAUM, P. S., BALDWIN, A., AND SAMSKY, J. On the useability and learnability of a transformationally generated subset of English. (In preparation)