[From: *Linguistics* 55 (1969)]

# TRANSLATION MODEL WITH SEMANTIC CAPABILITY

**L. W. TOSH**

Automatic translation systems may be characterized for the most part in one of two ways:

(1) Translators in which the language processors are internally dependent on the description of the syntax of the language being processed.

(2) Translators externally dependent on the syntax.*

Programming a translation system of the first type has two major advantages. It can be set up quicker and it requires less initial technical investment. On the other hand, such a system imposes many operational difficulties on debugging the linguistic description.

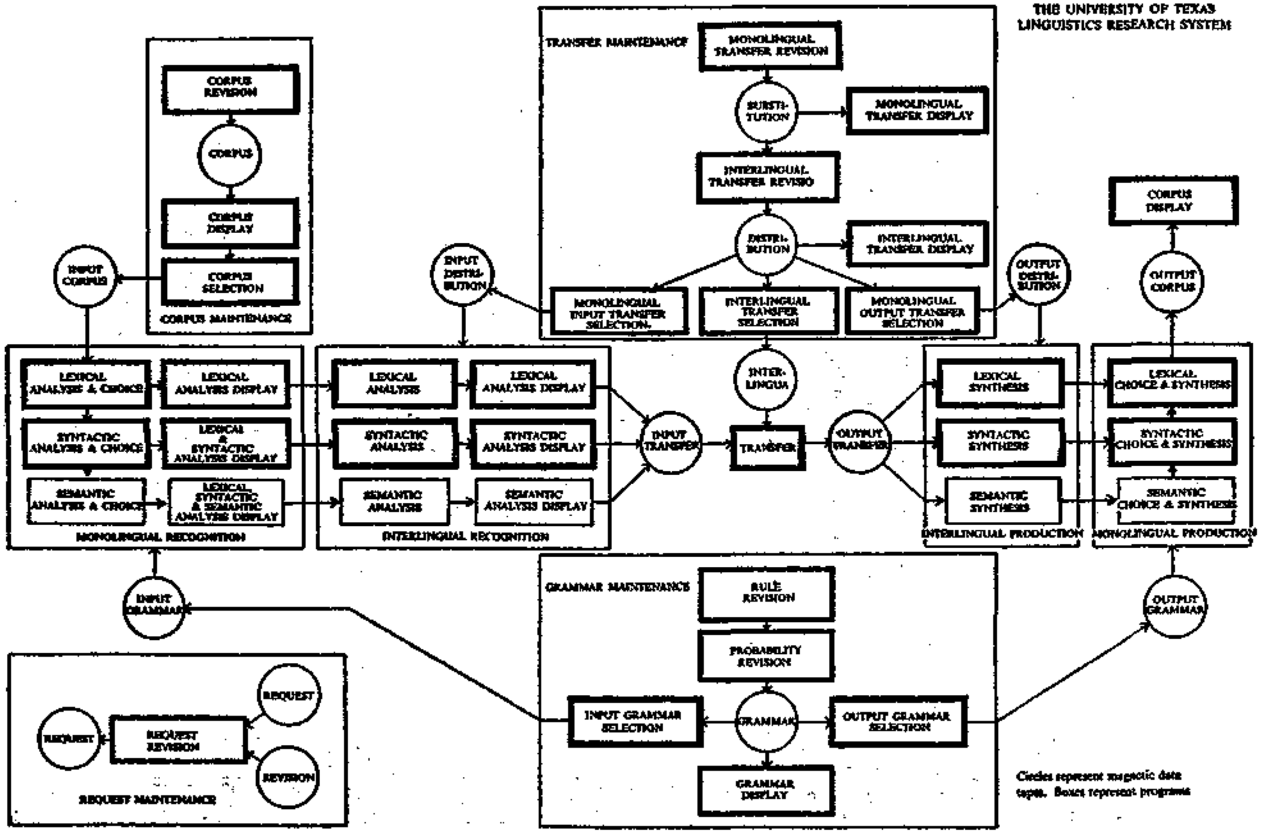
The second type is better adapted to more flexible linguistics research but is more costly and time-consuming to establish since its concrete results and more sophisticated research power and processing capabilities are not realized for some time after the initial effort.

The linguistic descriptive effort which has gone into either of these approaches has tended to concentrate primarily on matters of morphosyntactic analysis, rearrangement and synthesis. Such efforts as have been directed to semantic description in operational translation systems have been limited virtually to ad hoc statements of co-occurrence of particular items. Little else could have been done due, perhaps, to the lack of an adequate theoretic and formalizable notion of semantic process and to the descriptive limitations inherent in processing algorithms developed thus far.

In this paper I present an outline of the Linguistics Research System, a stratified linguistic data processing system, and interpret some of its facilities for application to semantic description. The system is of the second type in that the linguistic description is in the form of a phrase

* Support for the research was provided in part by the National Science Foundation, United States Army Electronics Command, and United States Air Force Systems Command.

THE UNIVERSITY OF TEXAS
LINGUISTICS RESEARCH SYSTEM



Circles represent magnetic data tapes. Boxes represent programs structure grammar with transformational capability and is independent of the underlying language processors.

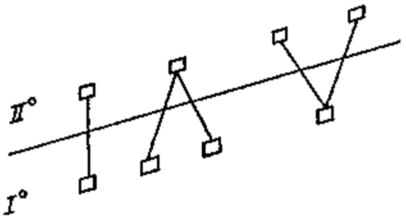
I shall discuss some of the linguistic relationships between two of the strata: the syntactic and the semantic. In order to avoid difficulties of terminology created by varying interpretations of the term SEMANTIC, I will often replace the pair of terms SYNTACTIC and SEMANTIC with a non-controversial pair, namely FIRST ORDER and SECOND ORDER. In describing our research, I will refer to features of Joos' semological model, Katz and Fodor's semantic model, and some elements of tagmemic theory. In addition, I have drawn on the taxonomic structure of Roget's *Thesaurus* to suggest that even his traditional categorization of language may lend itself to computational treatment. Finally, I try to show that the second order model is capable of accounting for some problems in transformation and discontinuity.

Let us consider now some general assumptions and features associated with the models in the Linguistics Research System. The restrictions imposed on grammatical descriptions within the framework of the models under study are not the same as in the case of generative grammars designed for random applications since the models are oriented toward machine translation. For instance, recursive use of symbols in rules is possible because the output of these grammars is not produced randomly. On the contrary, the output of our grammars is non-random because the input or source language to be translated consists of non-random sequences. Details of the analysis process are presented elsewhere (11; 12, pp. 19-34).

The flowchart, "Linguistics Research System", presents an outline of the stratified system of programs with which we are experimenting. I shall review briefly some features of the system, details of which are presented in (5).

A fundamental assumption in the analysis process (see MONOLINGUAL RECOGNITION on the flowchart) is that at the first order of analysis all parsings developed within that order of analysis are by definition well-formed. All of the parsing information thus obtained is carried forward to the second order of analysis. At the second order, however, some of the first order parsings may fail to be recognized and are thus by definition not well-formed with respect to the second order. We may say that some syntactic analyses will not be recognized as having meaningful interpretation. The basic problem then is to verify the well-formedness of parsings of any given order by submitting them to analysis at the next higher order.

In a stratified system of description, one may define various relationships as holding between any two orders of description. As Lamb pointed out in his view on stratificational theory, the following general relationships exist between what we have called first order parsing and second order parsings:



We may interpret the diagram as follows: in some instances there is a one-to-one correspondence between syntactic parsing and semantic interpretation. Then there are instances of different syntactic parsings having the same semantic interpretation. And finally, there are instances of a syntactic parsing which has more than one semantic interpretation. An example of a one-to-one correspondence of first and second order parsings would be the correspondence between the first order rule.

$$N_x \rightarrow \text{atomic weight}$$

and the second order rule

$$378.4.1 \rightarrow [N_x \rightarrow \text{atomic weight}]$$

where $N_x$ denotes a particular paradigmatic noun class and the decimal number 378.4.1 denotes a semantic or conceptual class derived from the taxonomic scheme found in Roget (cf. 8, p. 238). We can claim a one-to-one correspondence between the respective members of the classes $N_x$ and 378.4.1 based on the assumption that technical expressions like *atomic weight* have no synonyms.

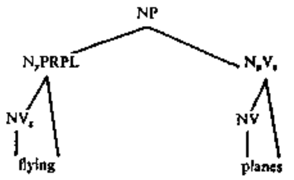
A simple example of two different first order parsings with the same second order interpretation can be found in first order rules as

$$ADJ_x \rightarrow AJ_y + \text{er } \textit{(pretti-er)}$$
$$ADJ_x \rightarrow \text{more} + AJ_z \textit{ (more beautiful)}$$

which will be members of the second order class

$$\text{COMPARATIVE} \rightarrow \left\{ \begin{matrix} ADJ_x \rightarrow A_y + \text{er} \\ \\ ADJ_x \rightarrow \text{more} + AJ_z \end{matrix} \right\}$$

The converse relationship is found in substrings like *flying planes* from the often quoted problem sentence *flying planes can be dangerous.* Depending on how we have designed our phrase structure grammar, we might obtain just one p-marker for the phrase, thus



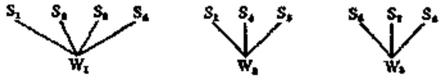
A second order classification of the above rule

$$NP \rightarrow N_yPRPL + N_PV_S$$

would have to provide for establishing at least two different transformational relationships for related strings like *to fly planes* and *planes which fly.* We may establish the transformational equivalences thus with second order rules like

$$\text{NO/CLS} \rightarrow \begin{Bmatrix} NP \rightarrow N_yPRPL^j \rightarrow N_pV_s{}^k \\ N_p \rightarrow N_pV_s{}^k + CLSREL^j \end{Bmatrix}$$
$$\text{INF/OBJ} \rightarrow \begin{Bmatrix} NP \rightarrow N_yPRPL^j + N_pV_s{}^k \\ INFPHR \rightarrow \text{to} + NV_x{}^j + N_pV_s{}^k \end{Bmatrix}$$

The symbol NO/CLS denotes a second order classification corresponding to the tagmeme noun + clause. The symbol INF/OBJ denotes the tagmeme infinitive + object. Superscripts *j* and *k* have been introduced to establish the sameness of semantic information over corresponding syntactic slots. $N_yPRPL^j$ denotes the first order classification of *flying* which is transformationally equivalent to $CLSREL^j$ or *which fly.* The first pair of second order rules above thus correlate *flying planes* to *planes which fly,* while the second pair *correlates flying planes* to *to fly planes.* In developing a model for second order descriptions, our task will be to provide a facility to account for phenomena such as Joos has described in his treatment of semology (2). Simply put, Joos has argued the notion that a word out of context is not without meaning but on the contrary is to be interpreted as having a maximum number of meaning values associated with it. The problem then becomes one of providing a means of selecting out the combination of values over a string of words to discover whether there are any interpretable combinations of values. Thus, in the illustrations below

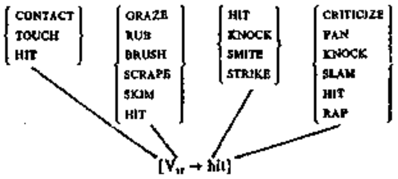


the problem is one of constructing a system to select out the predefined combination of semantic units (S) assigned over the string of words or morphemes ($W_k$).

The descriptive model with which we will work is designed to perform with rules of concatenation which select out such combinations of units. For instance, the expression *hit* might be classified in a first order rule thus

$$V_{tr} \rightarrow \text{hit}$$

The first order rule in turn might be multiply classified in the second order description,



where the set of symbols in each pair of braces may be taken to mark a distinctive semantic feature of *hit* and the set of braces marks the total range of meanings of *hit.* The example is intended only as an illustrative organization of data and is not complete.

If first order terminal rules are thus multiply classified in the second order description, we may proceed to write second order rules specifying the appropriate concatenation of semantic markers. Examples of such rules are presented below. Since the kind of rules presented here resemble those suggested by Katz and Fodor (3), let us consider first some of the features of their model.

In their discussion of semantic structure, Katz and Fodor propose a theory of structural parsing of semantic information. They give semantic markers for several examples, among them the sentence, *The man hit the colorful ball.* Figure 1 is an interpretation of one of the semantic parsings which they provide.

I have omitted the P-marker representation of the sentence and have represented the parsing instead in the form of a list structure for the sake of clarity. Each terminal expression is marked off by a box above which is a non-terminal symbol naming each respective list of terminal expressions. Thus, the expression *the* is a member of the list named T. Similarly, the expression *man* is a member of the list $N_c$. The combination of lists

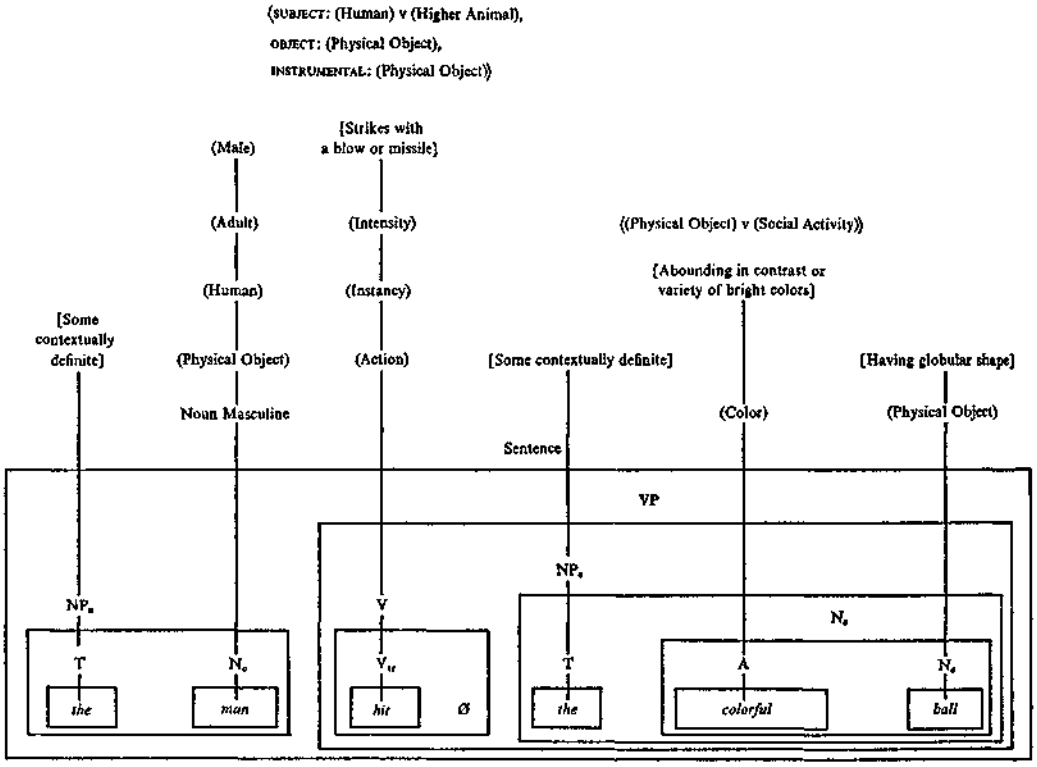

(SUBJECT: (Human) v (Higher Animal),
OBJECT: (Physical Object),
INSTRUMENTAL: (Physical Object))
[Strikes with a blow or missile]
(Male)
(Adult)
(Intensity)
((Physical Object) v (Social Activity))
[Abounding in contrast or variety of bright colors]
(Human)
(Instancy)
[Some contextually definite]
(Physical Object)
(Action)
[Some contextually definite]
[Having globular shape]
Noun Masculine
(Color)
(Physical Object)
Sentence
VP
NP_c
NP_c
V
N_c
T
N_c
V_tr
T
A
N_c
the
man
hit
Ø
the
colorful
ball

Fig. 1

T + $N_c$ constitute a complex member of the list $NP_c$. The usual branching diagram is thus represented by an equivalent list structure parsing.

Katz and Fodor provide a higher order parsing of each of the terminal items in such a way as to associate semantic information with each terminal. This information is represented by the vertical branches of Figure 1. The taxonomy in the illustration is the same as used in their article. If we now imagine a higher level branching system connecting the nodes given by Katz and Fodor, we have a complete parsing system similar to that of the syntactic system of P-markers, but of a higher order. An informal interpretation of such a higher order parsing system is given elsewhere (12, pp. 67-83). I have represented the system of Katz and Fodor in this manner in order to draw a parallel with the kind of second order system experimented with at the Linguistics Research Center.

Katz and Fodor make no claims for the adequacy of the semantic

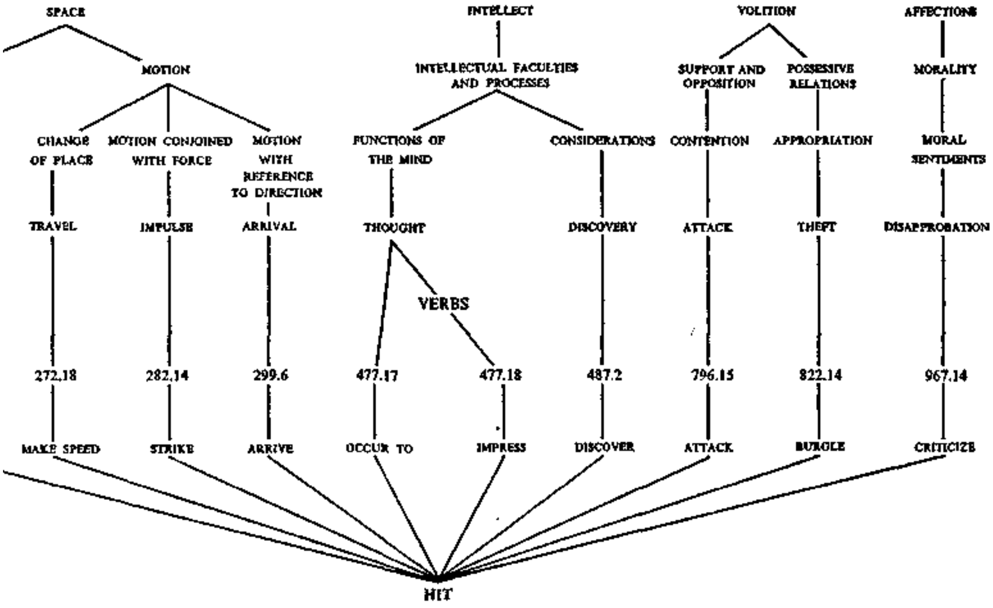


Fig. 2

taxonomy suggested in their illustrations. The examples are given merely to suggest an interpretation of their system. If we turn to some of the traditional sources of semantic classification such as Roget (8), we will find, for instance, that the classification of any one of the terminal strings in our illustrative sentence is considerably more detailed than might be assumed at first glance. The diagram in Figure 2 traces multiple classifications of the verb *hit* through Roget. The taxonomy in the illustration is derived directly from Roget. The verb falls into four principle classes: Space, Intellect, Volition, and Affections.

Such a set of distinctions can be incorporated into a second order parsing (Figure 3). Again, the first order parsing is represented in the form of a list structure diagram.

The second order parsing appears as the branching diagram superimposed on the first order parsing. Taxonomy in this example is derived from Roget. Symbols in single braces represent sets of synonyms or terminal items in the first order. Symbols in double braces are provided for the classification of first order terminal and non-terminal entries which Roget does not treat in his classification system.

The first order rule

$$V_{tr} \rightarrow \text{hit}$$

is a member of the second order class

$$\begin{Bmatrix}\text{HIT}\\ \text{KNOCK}\\ \text{SMITE}\\ \text{STRIKE}\end{Bmatrix} \rightarrow \boxed{V_{tr} \rightarrow \text{hit}}$$

The class also contains the members

$$\begin{Bmatrix}\text{HIT}\\ \text{KNOCK}\\ \text{SMITE}\\ \text{STRIKE}\end{Bmatrix} \rightarrow \boxed{V_{tr} \rightarrow \begin{Bmatrix}\text{knock}\\ \text{smite}\\ \text{strike}\end{Bmatrix}}$$

The non-terminal rule

$$V \rightarrow V_{tr} + \emptyset$$

which adds a zero affix to form the past tense is a member of the class

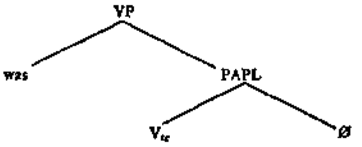
$$\{\{\text{REMOTE}\}\} \rightarrow \boxed{V \rightarrow \begin{Bmatrix}V_{tr} + \emptyset\\ V_{tr} + \text{ed}\end{Bmatrix}}$$

as is the rule which adds the affix *-ed.*

Second order non-terminal classes provide not only the capability of substituting synonyms but also the facility of transformations. The subtree in Figure 4 represents a second order parsing of the verb *hit* in the simple past tense, active voice. This second order subtree could be replaced by another as in Figure 5. The second order terminal class {{REMOTE}} will generate the subtree

$$\{\{\text{REMOTE}\}\} \rightarrow \boxed{VP \rightarrow V_{tr} + \emptyset}$$

or (as in Figure 5)



The second order terminal class {{BY}} will generate the first order rule which generates the preposition *by,* thus completing the passive counterpart of the active verb system. A complete parsing of the passive counterpart of the active sentence is shown in Figure 6.

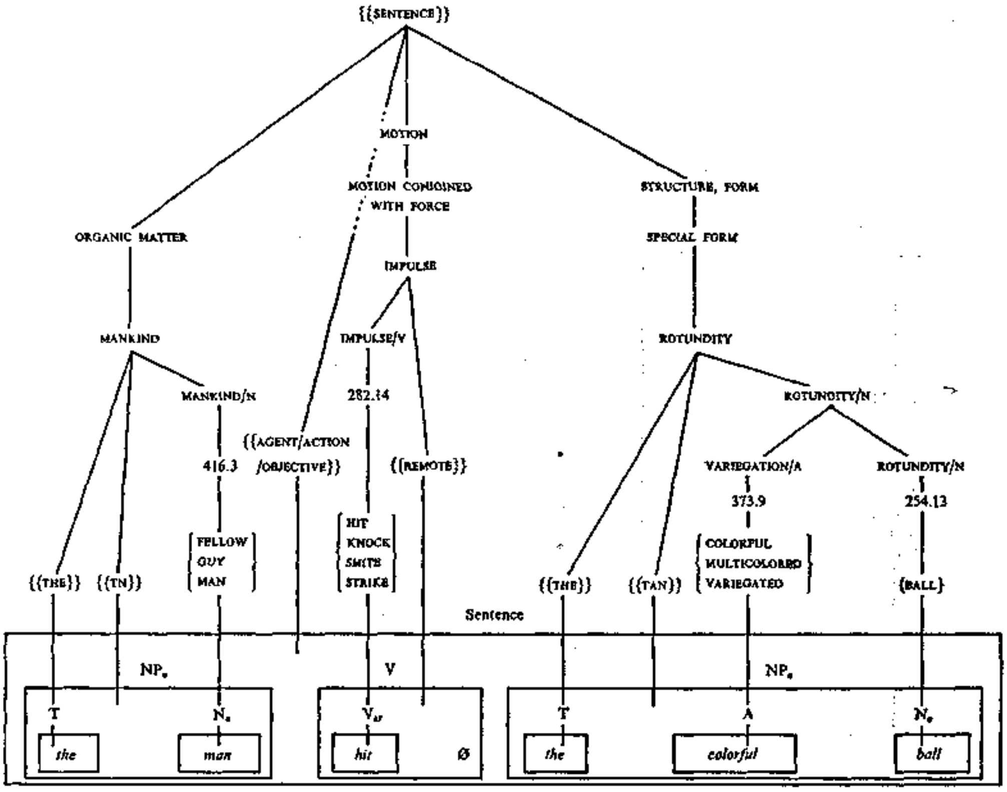


Fig. 3

In addition to synonyms and transformations, the second order description also provides for treatment of discontinuous constituents. In Figure 6 we can see that the first order parsing of the string *was hit by the man* establishes two phrase-level units: VP over *was hit* and ADV over *by the man.* Even though the passive markers *was ... ed* plus *by* are contiguous in this example, instances can be presented to show that other syntactic elements may intervene. Thus, while we may have defined *by* as a syntactic element not necessarily contiguous with the verb system, the relatedness of *by* with the verb system in expressing the passive agent is shown in the second order rule

IMPULSE → {{REMOTE}} + IMPULSE/V + {{BY}}

(Figure 6). This rule, when expanded by the rules

{{REMOTE}} → [VP → was + $V_{tr}$ + Ø]
{{BY}} → [P → by]

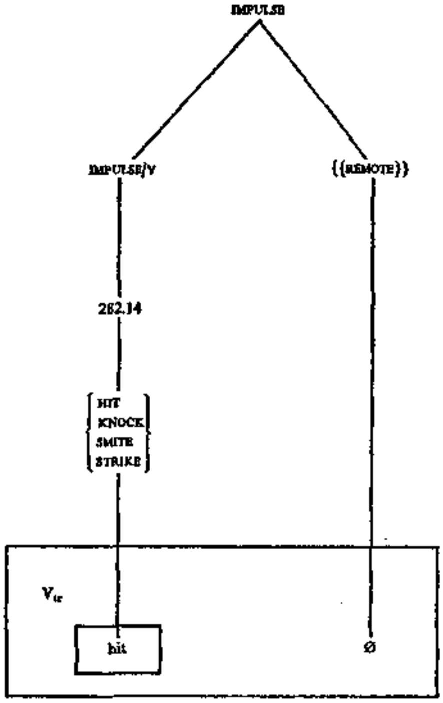


Fig. 4

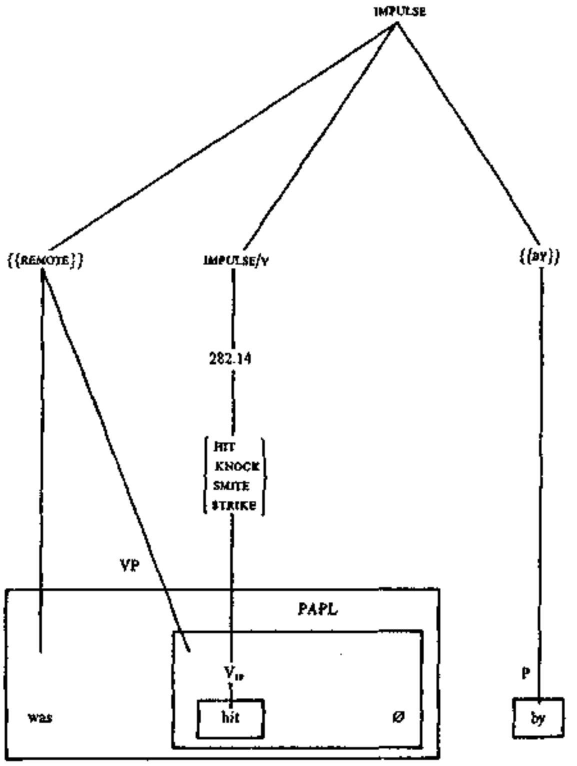
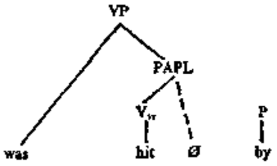
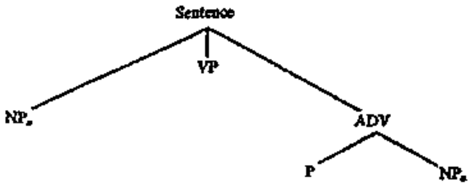


Fig. 5

will generate the first order elements



which are later concatenated by the first order subtree.



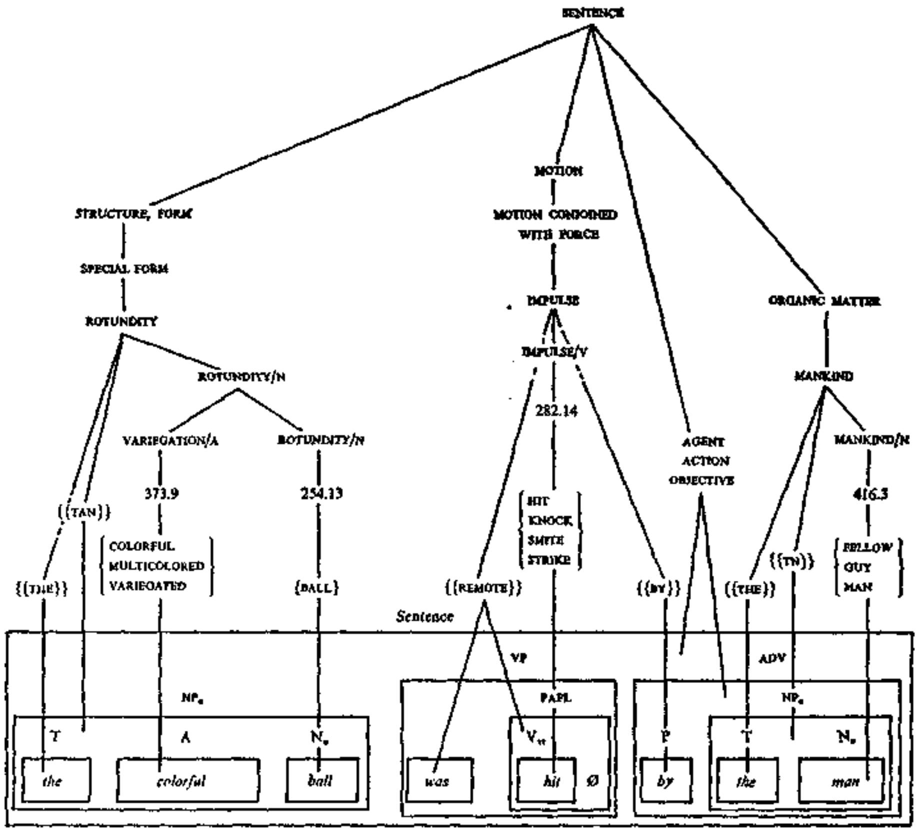


Fig. 6

Programs for translating at the second order were made operational early in 1966. We subsequently prepared linguistic data for German and English to test the operational status of the programs. The data were prepared for a few sentences only and on an ad hoc basis, which is to say that we were not concerned with the generality of descriptions. Taxonomy, for instance, was assigned in an arbitrary manner to expedite coding and does not represent interesting linguistic features.

The data were adequate, however, to provide satisfactory systems tests and to indicate the complexity to be anticipated in coding semantic descriptions. Table 1 is a sample of input-output data from the tests conducted on the programs. The German sentence is one of several input test samples. The English translations resulted from the set of linguistic data designed to provide some paraphrases.

TABLE 1

20,488<br>
WENN DIE MONDSCHEIBE DIE SONNE GANZ VERDECKT, ERSCHEINT EIN ROTER,<br>
10 — 15 BOGENSEKUNDEN BREITER RING UM DIE SONNE.

20488001 WHEN THE LUNAR DISK HIDES THE SUN COMPLETELY, A RED RING 10 TO<br>
20488001 15 SECONDS OF ARC IN WIDTH APPEARS AROUND THE SUN.

20488001 WHEN THE MOON’S DISK HIDES THE SUN COMPLETELY, A RED RING 10 TO<br>
20488002 15 SECONDS OF ARC IN WIDTH APPEARS AROUND THE SUN.

20488001 WHEN THE DISK OF THE MOON HIDES THE SUN COMPLETELY, A RED RING<br>
20488002 10 TO 15 SECONDS OF ARC IN WIDTH APPEARS AROUND THE SUN.

20488001 WHEN THE LUNAR DISK COMPLETELY HIDES THE SUN, A RED RING 10 TO<br>
20488002 15 SECONDS OF ARC IN WIDTH APPEARS AROUND THE SUN.

20488001 WHEN THE MOON'S DISK COMPLETELY HIDES THE SUN, A RED RING 10 TO<br>
20488002 15 SECONDS OF ARC IN WIDTH APPEARS AROUND THE SUN.

20488001 WHEN THE DISK OF THE MOON COMPLETELY HIDES THE SUN, A RED RING<br>
20488002 10 TO 15 SECONDS OF ARC IN WIDTH APPEARS AROUND THE SUN.

Since the time this article was submitted to press, research and development emphasis continued in the area of first order description of English, German and Russian. We also developed a Chinese-English lexicographic data base. Details of research are reported elsewhere (17, 18).

I am indebted to H. K. Ulatowska for reading the original presentation and suggesting improvements. This paper is based on a presentation made before the 1966 meeting of the Association for Machine Translation and Computational Linguistics.

*University of Texas at Austin*
*Linguistics Research Center*

## REFERENCES

(1) W. B. Estes, W. A. Holley, E. D. Pendergraft, “Formation and Transformation Structures”, LRC 63 WTM-3, Austin, Linguistics Research Center, May 1963.

(2) M. Joos, “Semology: A Linguistic Theory of Meaning”, *Studies in Linguistics,* 13 (1958), 53-70.

(3) J. J. Katz and J. A. Fodor, "The Structure of a Semantic Theory", *The Structure of Language* (Englewood Cliffs, 1964), 479-518.

(4) S. M. Lamb, *Outline of Stratificational Grammar* (Berkeley, 1962).

(5) W. P. Lehmann, "Symposium on the Current Status of Research", LRC 63 SR-1, Austin, Linguistics Research Center, October 1963.

(6) W. P. Lehmann, "Thirteenth Quarterly Progress Report", LRC 62 P-13, Austin, Linguistics Research Center, July 1962.

(7) E. D. Pendergraft, "A Generalized Computer System for Language Translation", LRC 64 WA-1, Austin, Linguistics Research Center, March 1964.

(8) P. M. Roget, *International Thesaurus* (3rd Edition, New York, 1962).

(9) D. A. Senechalle, "!Introduction to Formation Structures", LRC 63 WTM-2, Austin, Linguistics Research Center, April 1963.

(10) D. A. Senechalle, "Q-Collections and Concatenation", LRC 63 WTM-1, Austin, Linguistics Research Center, January 1963.

(11) L. W. Tosh, "Development of Automatic Grammars", *Linguistics,* 12 (March 1965), 50-60.

(12) L. W. Tosh, *Syntactic Translation* (The Hague, Mouton, 1965).

(13) L. W. Tosh, "System Requests Programming Reference Manual", LRC 65 TP-1, Austin, Linguistics Research Center, February 1965.

(14) R. Jonas, "System Design for Computational Linguistics", LRC 67 WA-1, Austin, Linguistics Research Center, in press.

(15) W. P. Lehmann and L. W. Tosh, "Research in German-English Mechanical Translation", Technical Report RADC-TR-67-98, Griffiss A. F. B., Rome Air Development Center, April 1967. .

(16) W. P. Lehmann, "Research on Syntactic and Semantic Analysis for Mechanical Translation", Final Report LRC 67 NSF 29, Austin, Linguistics Research Center, April 1967.

(17) W. P. Lehmann and L. W. Tosh, "Research in Chinese Lexicography", Technical Report RADC-TR-68-323, Griffiss A. F. B., Rome Air Development Center, November 1968.

(18) W. P. Lehmann, L. W. Tosh, R. R. MacDonald and M. Zarechnak, "Research in Russian-English Machine Translation on Syntactic Level", Technical Report RADC-TR-68-618, 2 vols., Griffiss A. F. B., Rome Air Development Center, March 1969.