[Proceedings International Congress of Linguists, Bucurest, 28 August – 2 September 1967]

L. W. TOSH

STRATIFICATIONAL GRAMMAR AND INTERLINGUAL MAPPING FOR AUTOMATIC TRANSLATION

Since 1959, the University of Texas Linguistics Research Center has been engaged in investigation into the possibilities of machine translation of languages. This has been a long-term project involving a number of funding sources. The effort it currently supported by contracts from the U.S. Air Force.

During this time, we have concentrated on the problems of syntactic description and the development of generalized parsing algorithms as tools for verifying descriptive research. Here, I propose to outline some aspects of the recent linguistic research. I will not deal with the technical aspects of the computer programming involved.

From the linguistic point of view there are two major areas of endeavor : one is pilot research of problematic areas; the second is comprehensive description deriving from the results of pilot research.

The principal feature which both types of effort have in common is the determination of the parameters of interlingual matching — parameters which become the foundation for subsequent structural description within a language.

The first step in the descriptive process is to determine the corresponding parts to be matched between pairs of sentences. For pilot descriptions we compare real pairs of sentences from actual texts. This channels the descriptive effort in the most productive direction toward our purpose : it focuses attention on the constructions characteristic of the particular area of scientific discourse with which we are dealing. For later comprehensive description, abstract classes of constructions on all levels are considered as well. Both traditional and more recent descriptive literature is consulted in all phases of research.

The initial determination of corresponding parts is done by submitting a pair of sentences to the judgment of bilingual informants. They specify the corresponding elements, providing us with the interlingual matching specifications. We then prepare a surface structure description within the framework set by these interlingual matching specifications.

Let me give you an example of our present and projected activity, using a pair of sentences from Russian and English. These demonstrate

some of the problems of matching various kinds of structures. The sentences are given in item (0) of the annex. The matching of lexical items is obvious so that we can write the necessary rules, such as, items (1), (2), and (3). The symbols N_k, N_1, etc. denote form classes of nouns. The classes are determined mainly by paradigmatic features, but extraparadigmatic, that is, syntagmatic, features are also considered. The symbols T_a, T_b, ..., T_z denote interlingual transfer classes. The transfer classes establish the correspondence between entries in pairs of monolingual grammars.

Morphological relationships must also be accounted for. They occasion such entries in the rule inventory as in item (4). Symbols NO_{FN} and $NOsing_{AN}$ denote classes of inflected nouns. The linking symbol *T,* denotes the interlingual transfer class. It contains all inflecting entries for the singular. For a comprehensive description of a language, it is necessary to write the entire set of such entries into the inventory of rules. This would have to cover all instances of morphological inflection — for all adjectives, nouns, verbs, and any other inflecting classes.

Moreover, some constructions have greater complexity. In these sample sentences, the noun phrases exemplified by этой окрестности or *(of) this area,* for instance. Their surface structures are to be accounted for by items (5), (6), and (7). Here for the Russian, the symbol NP_{FG} denotes the noun phrase construction feminine and genitive. Similarly, the symbol *DT* denotes determiners or articles and the symbol *NO,* the classification of inflected nouns. Here, этой is classified as a determiner *(DT).* Full subscripts, such as *FGDPI,* thus indicate the gender and the range of cases which characterize the form or construction so classified. In this instance, the gender is feminine and the case range would be genitive, dative, prepositional, and instrumental.

Then, in a similar fashion, corresponding structures are established for the English counterpart. For this, there may be such entries as items (8) and (9). Here, the symbol NP_{sinq} classifies the noun phrase construction as singular. Symbols like $NO_{sinq.A}$ denote a further classification of inflected nouns. They are used to include some extraparadigmatic features as cooccurrence with the indefinite article *a* or *an.*

If we look at a tree model of the description outlined here, we will see an analysis like item (10) for the Russian noun phrase and item (11) for the English noun phrase. You can see that the two parsings do not match element for element.

Nevertheless, the nodes marked *NP* at the top correspond for the phrases which are subsumed by these nodes. Likewise, the nodes DT_{FGDPI} and *DT* correspond for the expressions этой and *this.* Finally, there is correspondence between the symbols NO_{FGDP} and $NO_{sinq.AN}$ for окрестности and *area,* respectively.

What we now need is to establish interlingual transfer correspondence between such differing but equivalent nodes in a pair of parsing structures.

It is in such areas that recent efforts at the Linguistics Research Center have been focused. We have developed parsing algorithms of a very wide generality. Therefore, we are able to write transfer entries over these and more complex trees.

Such transfer entries would present the descriptions to the parsing algorithms in such a guise as item (12). In this example, the * denotes that the entries shown do not exist as explicit phrase structure rules in the respective grammars. Rather, they are implicitly defined in the interlingual transfer system. Such a mechanism defines the class of bi-directional transformations on the surface structures.

To further illustrate this feature, let us consider the problem of correlating discontinuous elements in a string. In the pair of example sentences there are two occurrences of discontinuities. One is in the verb phrase illustrated in item (13). The other is in the correlative conjunction structure in item (14). The surface structure entries devised for part of the structure over the two verb phrases are given in items (16) through (23). Interlingual transfer is shown by the classes T_k, T_l, and T_n items (15), (16) and (18). These transfer mappings are straightforward and they will generate the matrix structure into which the rest of the structures for the lexical discontinuity will fit. Transfer is shown later for items (17) and (19) through (23). In diagram form, the matrix structure for Russian would appear as in item (24). A corresponding structure would be generated in the output for English. The more complex problem remains, that of accounting for the discontinuities in the lexical string.

For this, the notational form (in our system of approach to such discontinuities) is similar to that given earlier for less complex problems. The list of entries pertaining to discontinuous elements is given in items (25) through (28). The symbol *ZAVIS.OT* denotes the class specifying the discontinuous concatenation of the rule given in item (17) with the class *OT*. If rule (28) were applied to rule (25) we would generate the structure in item (29).

The parsing algorithms which manipulate the descriptive rules may be regarded as composed of two main systems : primary and secondary. The primary algorithm manipulates entries of the kind shown in solid lines, for instance the structure in item (20). The secondary algorithm manipulates entries of the kind shown in broken lines, for instance the structure in item (31). In items (25) through (27) the symbol *d* denotes a function similar to + in phrase "structure rules. However, unlike +, *d* does not denote an immediate concatenation between symbols. Instead, *d* denotes that the two elements concerned are more loosely associated, that is, in a discontinuous manner. See for instance the association of solid structures in item (29).

We may generate the corresponding English structure similarly from the class *DEPEND.ON* as in item (32). By applying all of items (25) through (28) we may generate the more complex structure with the correlative conjunctions as in item (33). Careful inspection of item (33) will

reveal that each of the uppermost symbols over solid lines fits into the matrix structure in item (24). I have not illustrated the English counterpart, since it is structured quite similarly.

As I have mentioned, the relationship between primary and secondary algorithms has been kept purposely general. In this way, we can accommodate mappings or translation problems which may involve both primary and secondary algorithms in one language and only the primary algorithm in the other. An example of this type of problem is to be found in the translation of the noun phrase железной дороги into *railroad.* Here, the Russian input is regarded as discontinuous with respect to translation, since the lexical stems are interrupted by inflectional endings. Item (34) shows a parsing of the Russian string. By writing the set of transfer rules shown in items (35) and (36), we may map onto the English structure in item (37). The new notation <α, α >and < α > represents the interlingual mapping mechanism. It records the mapping equivalences of class symbols, between grammars. Thus, the notation in item (35) denotes that the lexical contents of the positions α and α in Russian are to be mapped into the single position α in English. It is also possible to perform the reverse operation : under the condition that the matrix structure has been defined, one content position may be mapped into many. Mapping notation was not used earlier because the relationships were straightforward. In actual practice, however, the relationship must be defined for each set of equivalent interlingual transfer entries.

There is one final problem area of interest in the examples, and that is the generation of articles in English. This was briefly touched upon earlier. I do not intend to suggest that the descriptive problem has been solved, but only that the parsing algorithms are geared to manage the descriptive data. In translating the phrase построения железной дороги we may generate several alternatives such as item (38). For the alternative the *building of a railroad,* items (39) and (40) list the transfer entries necessary for mapping the Russian and English. Item (39) will generate the Russian structure shown in item (41) and the English structure shown in item (42). When item (40) is applied it will fill in the missing inflectional step in both languages. The symbol *VO* in the secondary parsing denotes the class of constructions which will come to account for transformation between verb plus object and deverbalized noun plus object.

The symbol $N_{DE\text{-}VERB}$ denotes this class of deverbalized nouns. In English, such must be accompanied by some matrix of expression like *the ... of a ...* whenever the translation alternative with *of* is chosen.

For the present, the descriptive solutions offered here are but projections, since research into description utilizing secondary algorithms has only begun.

However, we have verified the performance of the primary algorithms with a reasonably large data base, one which is very comprehensive for English, German and Russian noun phrases (exclusive of relative clauses).

This data base contains dictionary entries numbering some 190,000, 40,000 and 140,000 items respectively. Entries describing syntactic constructions number several thousand in each language. Small, pilot descriptions of a few hundred entries have been tested in French, Spanish, Hebrew, Japanese, and Chinese. With the development of the secondary algorithms, we are beginning research into the description of problems involving discontinuity, transformations, and semantic features. In conclusion, one might say, "successful description of these areas will depend both on the developing of the secondary algorithm and on good fortune". I am indebted to my colleagues, Mrs. H-J. Hewitt and Mr. S. Whelan, for their aid in matters of presentation.

REFERENCES

R. W. JONAS, *Generalized Translation of Programming Languages,* AFIPS Conference Proceedings, 1967 (Fall joint Computer Conference, Vol. 29 (in- press).

R. W. JONAS, *System Design for Computational Linguistics* (Austin, 1967).

W. P. LEHMANN, *An Experiment in Machine Translation,* "The Graduate Journal", (1965) 7.111—131.

W. P. LEHMANN, *Research in German-English Mechanical Translation,* Technical Report RADC-TR-67-98 (Griffiss Air Force Base, 1967).

W. P. LEHMANN, *Research on Syntactic and Semantic Analyses for Mechanical Translation:* (Austin, 1967).

L. W. TOSH, *Content Recognition and the Production of Synonymous Expressions,* Proceedings of the ninth International Congress of Linguists (The Hague, 1964), p. 722—729.

L. W. TOSH, *Development of Automatic Grammars,* "Linguistics" (1965), 12.49—60.

L. W. TOSH, *Initial Results of Syntactic Translation at the Linguistics Research Center,*. "Linguistics" (in press).

L. W. TOSH, *Syntactic Translation* (The Hague, 1965).

L. W. TOSH, *Translation Model with Semantic Capability,* "Linguistics" (in press),

ANNEX

(0) **Индустриализация этой окрестности зависела и от построения железной дороги и от электрификации.**

Industrialization of this area depended both on the building of a railroad and on electrification.

(1) $[N_k \rightarrow$ **индустриализация**$] \leftarrow T_a \rightarrow [N_a \rightarrow$ **industrialization**$]$

(2) $[N_1 \rightarrow$ **окрестность**$] \leftarrow T_b \rightarrow [N_b \rightarrow$ **area**$]$

(3) $[N_s \rightarrow$ **ящур**$] \leftarrow T_z \rightarrow [N_z \rightarrow$ **hoof-and-mouth disease**$]$

(4) $$\begin{bmatrix} NO_{FN} \rightarrow \begin{Bmatrix} N_k + Я \\ N_y + А \end{Bmatrix} \\ NO_{FG} \rightarrow N_x + И \\ \vdots \end{bmatrix} \leftarrow T_i \rightarrow \begin{bmatrix} NO_{sing.AN} \rightarrow \begin{Bmatrix} N_a \\ N_b \\ N_c \end{Bmatrix} + \varnothing \\ \vdots \end{bmatrix}$$

(5) $$NP_{FG} \rightarrow DT_{FG} \rightarrow NO_{FG}$$

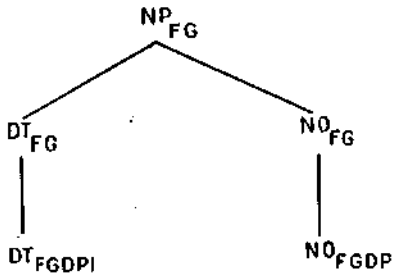
(6) $$DT_{FG} \rightarrow DT_{FGDPI}$$

(7) $$NO_{FG} \rightarrow \begin{Bmatrix} NO_{FGDP} \\ NO_{FGDPI} \\ \dots \end{Bmatrix}$$

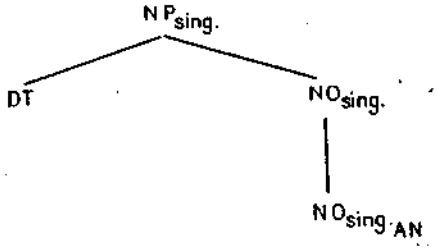
(8) $$NP_{sing.} \rightarrow \begin{Bmatrix} DT \\ DT_{sing.} \end{Bmatrix} + NO_{sing.}$$

(9) $$NO_{sing.} \rightarrow \begin{Bmatrix} NO_{sing.A} \\ NO_{sing.AN} \end{Bmatrix}$$

(10)



(11)



(12)* $$[NP_{FG} \rightarrow DT_{FGDPI} + NO_{FGDP}] \leftarrow T_j \rightarrow *[NP_{sing.} \rightarrow DT + NO_{sing.AN}]$$

(13) зависела ... от .... от
depended ... on ... on

(14) и ... и
both ... and

(15) $[VP_F \rightarrow VP_F \rightarrow ADV] \leftarrow T_k \rightarrow [VP \rightarrow VP \rightarrow ADV]$

(16) $[VP_F \rightarrow V_x + ела] \leftarrow T_1 \rightarrow [VP \rightarrow V_z + ed]$

(17) $V_x \rightarrow$ завис $\quad V_z \rightarrow$ depend

(18) $[ADV \rightarrow C + ADV + C + ADV] \rightarrow T_m \leftarrow [ADV \rightarrow C + ADV + + C + ADV]$

(19) $C \rightarrow$ и $\quad C \rightarrow$ both

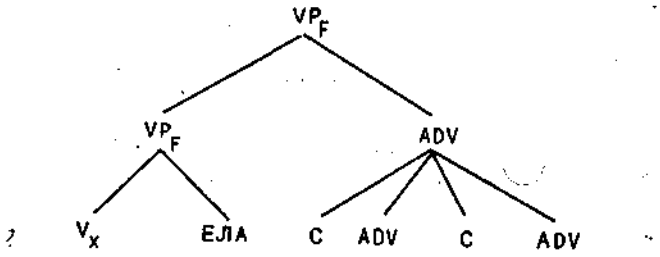
(20) $ADV \rightarrow PRPH \quad ADV \rightarrow PRPH$

(21) $PRPH \rightarrow PRP_G + OBJ_G \quad PRPH \rightarrow PRP + OBJ$

(22) $PRP_G \rightarrow$ ОТ $\quad PRP \rightarrow$ on

(23) $C \rightarrow$ and

(24)



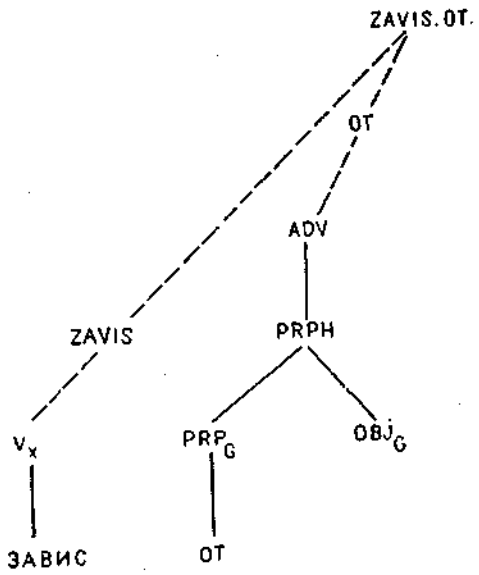
(25) *[ZAVIS. OT → [$V_x \rightarrow$ завис] d OT] ← T_n → [DEPEND. ON → [$V_y \rightarrow$ depend] d ON]

(26) [OT → CC d OT d OT] ← T_o → [ON → CC d ON d ON]

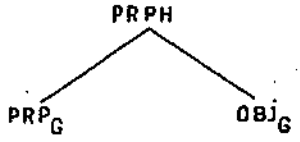
(27) [CC → [C → и] d [C → и]] ← T_p → [CC → [C → both] d [C → and]]

(28) [OT → *[ADV → OT + OBJ]] ← T_q → [ON → *[ADV → on + + OBJ]

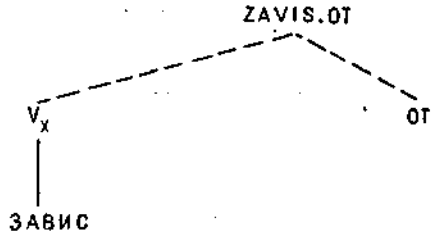
(29)


ZAVIS.OT
OT
ADV
ZAVIS
PRPH
V_x
PRP_G
OBJ_G
ЗАВИС
OT

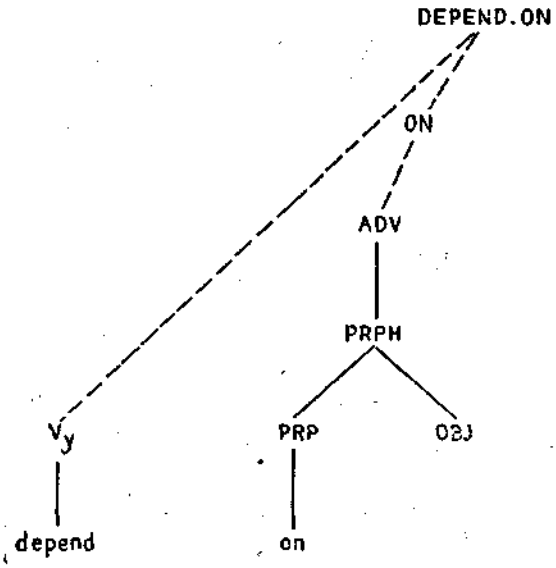
(30)


PRPH
PRP_G
OBJ_G

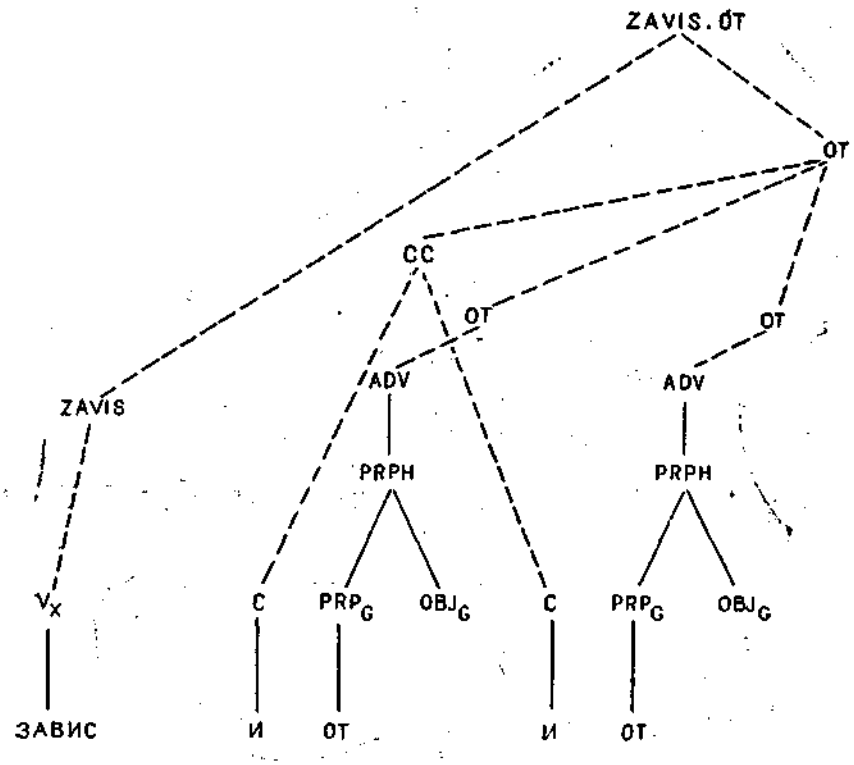
(31)


ZAVIS.OT
V_x
OT
ЗАВИС

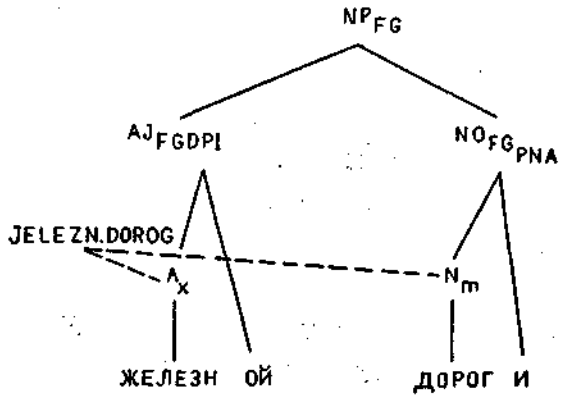
(32)


DEPEND.ON
ON
ADV
PRPH
V_y
PRP
OBJ
depend
on

(33)


ZAVIS.OT
OT
CC
OT
OT
ADV
ADV
ZAVIS
PRPH
PRPH
V_x
C
PRP_G
OBJ_G
C
PRP_G
OBJ_G
ЗАВИС
И
ОТ
И
ОТ

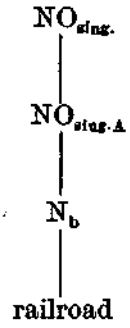
(34)



(35) $*[NP_{FG} \rightarrow \underset{<\alpha}{A_x} + \text{ой} + \underset{\alpha>}{N_m} + \text{и}] \leftarrow T_r \rightarrow *[NO_{sing.} \rightarrow \underset{<\alpha>}{N_b}]$

(36) $[JELEZN.DOROG \rightarrow [A_x \rightarrow \text{желэзн}]\ d\ [N_m \rightarrow \text{дорог}]] \leftarrow T_s \rightarrow [N_b \rightarrow \text{railroad}]$

(37)



(38) building $\left(\left\{\begin{matrix}\text{a}\\ \text{the}\end{matrix}\right\}\right)$ railroad (s)

the building of $\left(\left\{\begin{matrix}\text{a}\\ \text{the}\end{matrix}\right\}\right)$ railroad(s)

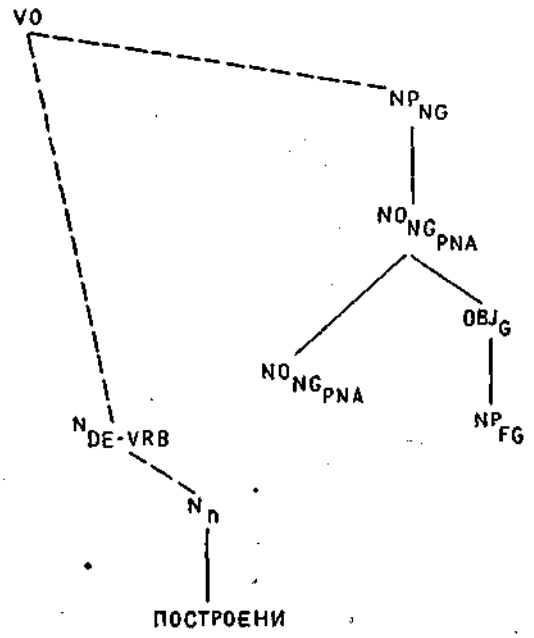
(39) $[VO \rightarrow *[NP_{NG} \rightarrow \underset{<\alpha}{NO_{NG_{PNA}}} \underset{\leftarrow T_t \rightarrow}{} + \underset{\beta>}{NP_{FG}}]\ d\ [N_n \rightarrow \text{построени}]]$

$[VO \rightarrow *[NP_{sing.} \rightarrow \text{the} + \underset{<\alpha}{NO_{sing.}} + \text{of a} + \underset{\beta>}{NO_{sing.}}]$

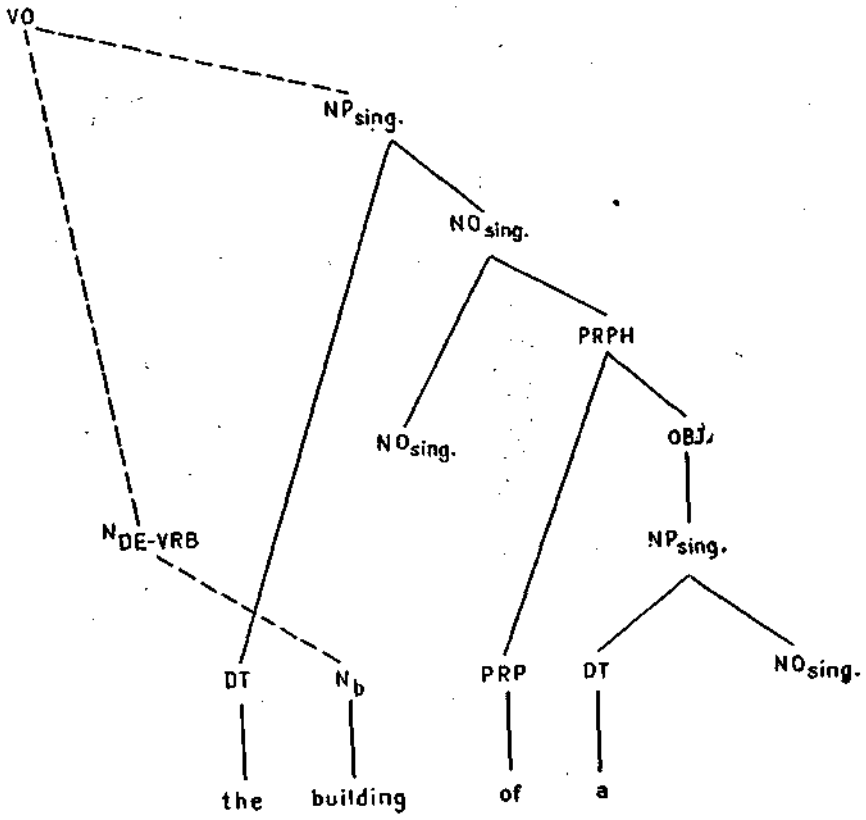
$d\ [N_b \rightarrow \text{building}]]$

(40) $[NO_{NG_{PNA}} \rightarrow N_n + \text{я}] \rightarrow T_i \rightarrow *[NO_{sing.} \rightarrow N_b]$

(41)


VO
NP NG
NO NG PNA
OBJ G
NO NG PNA
NP FG
N DE-VRB
N n
ПОСТРОЕНИ

(42)


VO
NP sing.
NO sing.
PRPH
NO sing.
OBJ
NP sing.
N DE-VRB
DT
N b
PRP
DT
NO sing.
the
building
of
a