[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 3: CURRENT RESEARCH

RESEARCH IN MACHINE TRANSLATION

Harry H. Josselson
Wayne State University

The Russian Machine Translation project, sponsored by the Office of Naval Research, has been carried out at Wayne State University since July 1958. Our group consists of a linguist, Prof. H.H. Josselson, Chairman, Department of Slavic Languages; Profs. A.W. Jacobson and Charles Briggs, Department of Mathematics; a senior programmer, Amelia Janiotis; several junior programmers, student assistants with a substantial command of Russian, and other punched-card clerical people. Prof. Paul Garvin of Georgetown University is serving as a consultant to the group and has greatly stimulated the direction of our thinking.

We had definite ideas as to the approach to this complex problem. We felt that the most effective attack would be one where the material was limited to a fairly narrow area in a subject-field. We were naturally concerned about doing repetitious work, going over the same things in a way that had already been done elsewhere. To avoid this overlapping of effort, both in regard to the method of attack and the areas of investigation, we have diligently searched the literature and have personally been in contact with most of the centers in which work is being done in this field. We are very pleased to have had the cordial and helpful cooperation wherever we have sought contact and explanation of the work that was being done.

On the basis of these discussions with other groups, we have become more assured that our basic approach to this problem is reasonable and effective. Likewise, we are now aware of what is being done here and abroad and can thus avoid directly repetitious work. It is, of course, natural that in the beginning a certain amount of ground must be covered which is common to all investigations.

In accordance with our decision to limit the subject-field, we have chosen an area in mathematics dealing with partial differential equations. This restriction will limit not only the vocabulary but also, we believe, the structural diversity inherent in the general

language. Accordingly, we have chosen three Russian articles, one from математический сборник , 1955, "The solution of problems of Cauchy for certain types of systems of linear partial differential equations" by V. M. Borok, second from успехи математических наук, 1953, "Fourier transforms of rapidly increasing functions and questions of the uniqueness of the solution of Cauchy' s problem" by I. M. Gel'fand and G. E. Shilov, and third from успехи математических наук, 1954, "On the solution of Cauchy' s problem for regular systems of linear partial differential equations" by A. G. Kostyuchenko and G. E. Shilov. For these there exist translations prepared by the American Mathematical Society. This gives us parallel texts to work with. Incidentally, we have made changes in the English translations in order to bring them closer to Russian forms while retaining them in acceptable English. We feel that the parallel-text approach has numerous advantages in relating the two languages structurally as well as in the specific resolution of ambiguities in meaning and form.

Another feature of our program is that we are aiming at a careful linguistic analysis of the material prior to any effort to program work for a computing machine. What we wish to say is that our main area of attention concerns the structural analysis of the language for the purposes of mechanical translation. We are not concentrating especially on the problems of glossaries; we simply confine our effort to such questions as multiple meaning, insertion, or deletion. We believe that these questions can be resolved only in the over-all analysis of structure, perhaps with the aid of semantic considerations. Our effort concerns the determination of clause boundaries and the isolation of other lexical groups which must be carried out if automatic procedures are to be arrived at for their translation into another language. To describe our attack in general then, we would say that (1) we are working with a small subject-field; (2) we are concentrating mainly on problems of ambiguity, both on the lexical as well as the morphological level, and of rearrangement, laying aside for the moment other equally important problems; and (3) our main procedure involves structural analysis and the use of parallel texts. With these rules as our guides, we aim at developing practical translation procedures yielding fluent and accurate text.

We believe that postediting in actual production translation work will be necessary, but, as more experience is gained and procedures are refined, the amount of postediting will diminish.

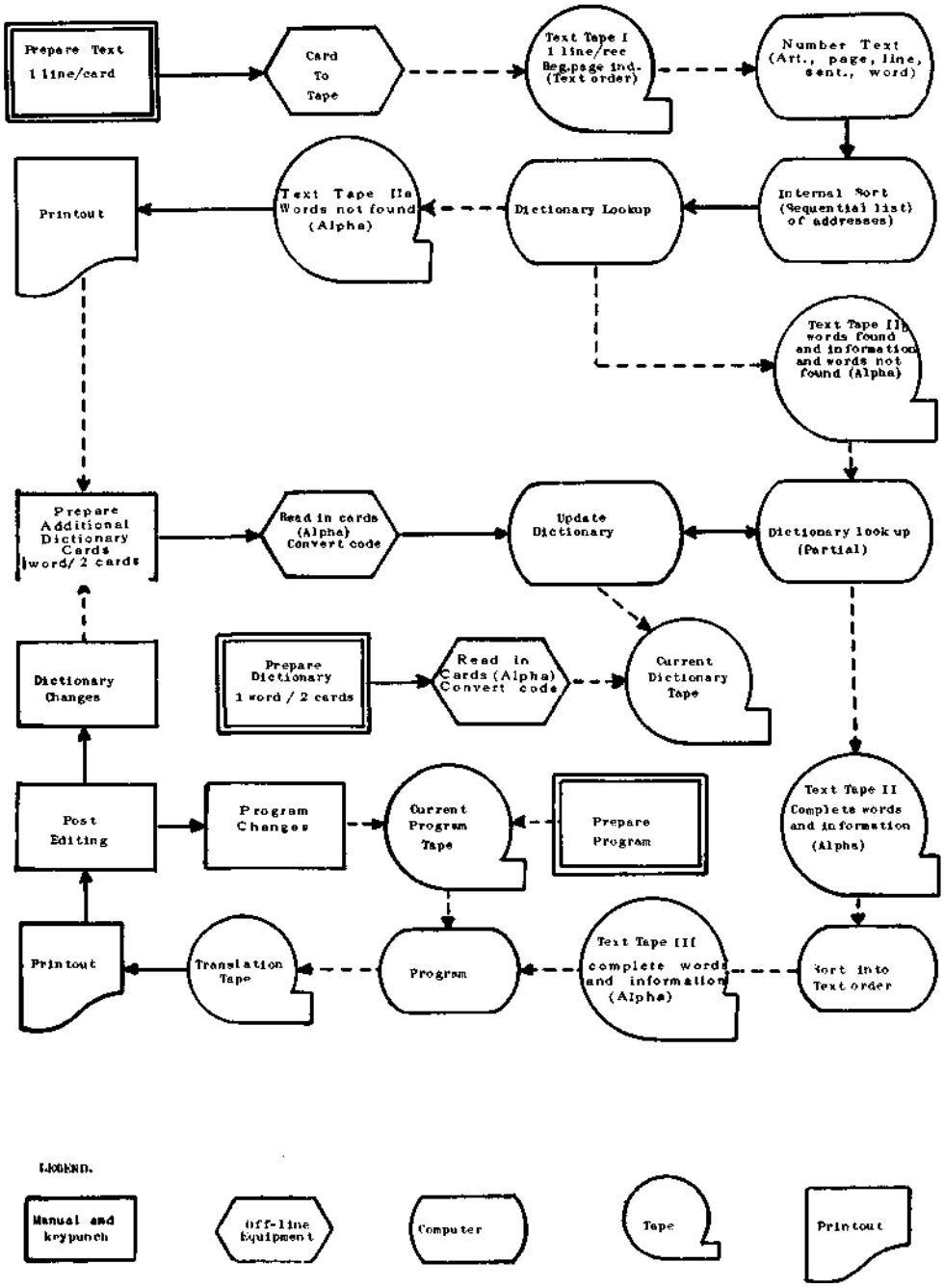
We will now describe our general procedure. It is characterized by three distinct starting points for processes which ultimately merge. These starting points are (1) preparation of the program, (2) preparation of the dictionary, and (3) preparation of the text on cards. Figure 1 represents an outline of our general procedure.

The input to the procedure loop involving the program consists of the current program tape and a text tape which comprises the words encountered in the text (in text order) and the dictionary information about each word. The output is a printout of the translation effected, which when subjected to a postediting procedure will be the basis for program changes. Thus, ultimately, a new current program tape will result.

The postediting may indicate the necessity for certain dictionary changes, as well as program changes, and this leads us to the dictionary-procedure loop. A dictionary based on words encountered in a number of texts (mathematical, in our case) *is* prepared on cards. The information includes the Russian word, grammatical and syntactic information, and the English translations of the word. These cards will be sorted into alphabetical order and converted to tape and, in the process, the abbreviated card codes will be expanded to the binary code called for in the program. Obviously, it will be continually necessary to add new entries to the dictionary, as well as to correct entries on the basis of experience gained through program runs. The additional entries necessary will be indicated by the failure to find words in the dictionary which appear in a new text being processed. Thus, there is a provision in the loop to update the dictionary and, simultaneously, to fill in the gaps in the text tape. Since the text will be in alphabetical order for the purpose, it must be resorted to text order for input into the program loop.

Finally, we note that the preparation of the text involves punching the text on cards (one line per card); converting to tape (one line per record); numbering each word (according to article, page, line, sentence, and occurrence in a sentence); sorting the words into alphabetical order; performing a dictionary lookup; and producing a tape

Figure 1 - OUTLINE OF GENERAL PROCEDURE



(which will be completed later) containing the words found in the dictionary with the information therein, together with the words not found, and a separate tape of the words not found, from which the aforementioned addition to the dictionary will be prepared.

Since some notable work has already been accomplished in the machine translation field by other groups, and since we have greatly profited therefrom, we are not concentrating on investigating extensive syntactic analysis routines in order to produce a better than word-for-word translation. Our work consists of testing some techniques of MT which, essentially, are experiments with sophisticated concepts of syntax conceived from the point of view of machine translation.

We are working with particular problems in syntax, using the computer to test our routines. We will be using an IBM 709 computer, with 32,000 words of storage and two data-synchronizer channels with two 12-tape units on line. This computer is located at Chrysler Corporation, Detroit, who have been very cooperative in allowing us to use their computer in our work. Our technique is to isolate a problem by manually simulating or by excluding all those parts of a translation program which are not directly concerned with the problem under consideration. The problems that we deal with are formulated in terms of the over-all syntax concepts of Paul Garvin, who is a consultant to our project, as mentioned earlier. The purpose of this part-by-part experimentation is to clarify the relationship between various rules and routines. As an example, governing-modifier packages are contained in prepositional structures and conversely. Hence, the rules about governing-modifier packages and prepositional structures must be brought into sensible relationship with one another. The homograph-resolution routines preceded all of the other routines in our first approach to the problem, but in view of certain revisions in some of the rules, we may find it more convenient to solve some of the homographs at the beginning of the program and others later on. An example of the technique of isolation is found in our present nominal blocking pass, which is explained below in this paper. An earlier Ramo-Wooldridge program, the truncated syntax program, made use of the concept of agreement checking between nouns and their modifiers in major syntax routines.

For greater generality, we now think that agreement checking should be realized in a separate pass, prior to the main syntax passes. The experiments with the truncated syntax program have indicated exactly how much should be included in a nominal block in order that it be maximally useful in the rest of the analysis. In view of this restructuring, noun-adjective ambiguities (e.g., данные = "given", "data") not picked up as adjectives in nominal blocks may turn out to be nouns. Also, predicate-adverb-preposition homographs may be resolved as adverbs. Thus, homograph rules may be re-written more efficiently and better sequenced, as well.

In order to test our routines on the computer, we have decided for the present to simulate the dictionary lookup, ignore the homograph problem (in certain instances), and ignore the English translation output problem in the initial stages of programming by limiting the output to an indication of block boundaries. At this point it is appropriate to mention that our ultimate goal is a sentence image whose elements are properly labeled blocks (with indication of syntactic function, e.g., subject, and of grammatical class membership, i.e., capacity for syntactic function) which can be manipulated by syntactic rules and related to semantic rules to produce high-quality translation.

As a result of syntactic analysis from the MT point of view, it has been necessary to depart from the traditional part-of-speech scheme and consider word classes and sub-classes in terms of syntactic function and distribution (e.g., adverbs which modify only those items to the right of them, like очень = "very"). These new form classes are presented in detail in Figure 2.

These considerations of syntactic function and distribution, in addition to purely morphological criteria, have resulted in a reshuffling of form classes of Russian used in traditional or even descriptive grammar. Thus, a major word class of conjunctions is divided into "ambiguous, non-ambiguous", "finite, infinite", "contains бы, does not contain бы". A new major word class "modifiers" cuts across traditional divisions by including adjectives, participles, numerals, demonstrative pronouns, and possessive pronouns. A comparison of the new and old grammar classes is contained in Figure 3.

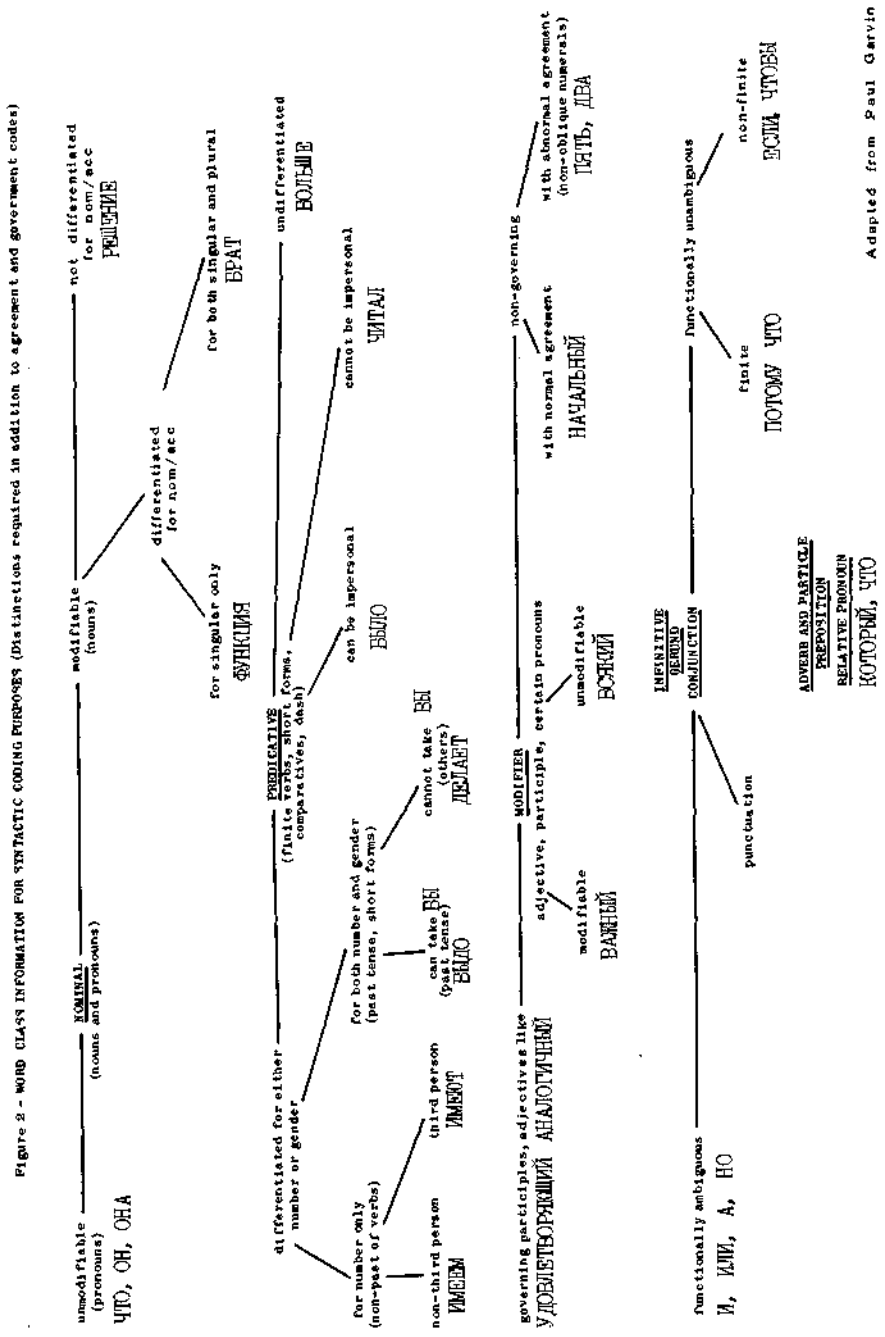
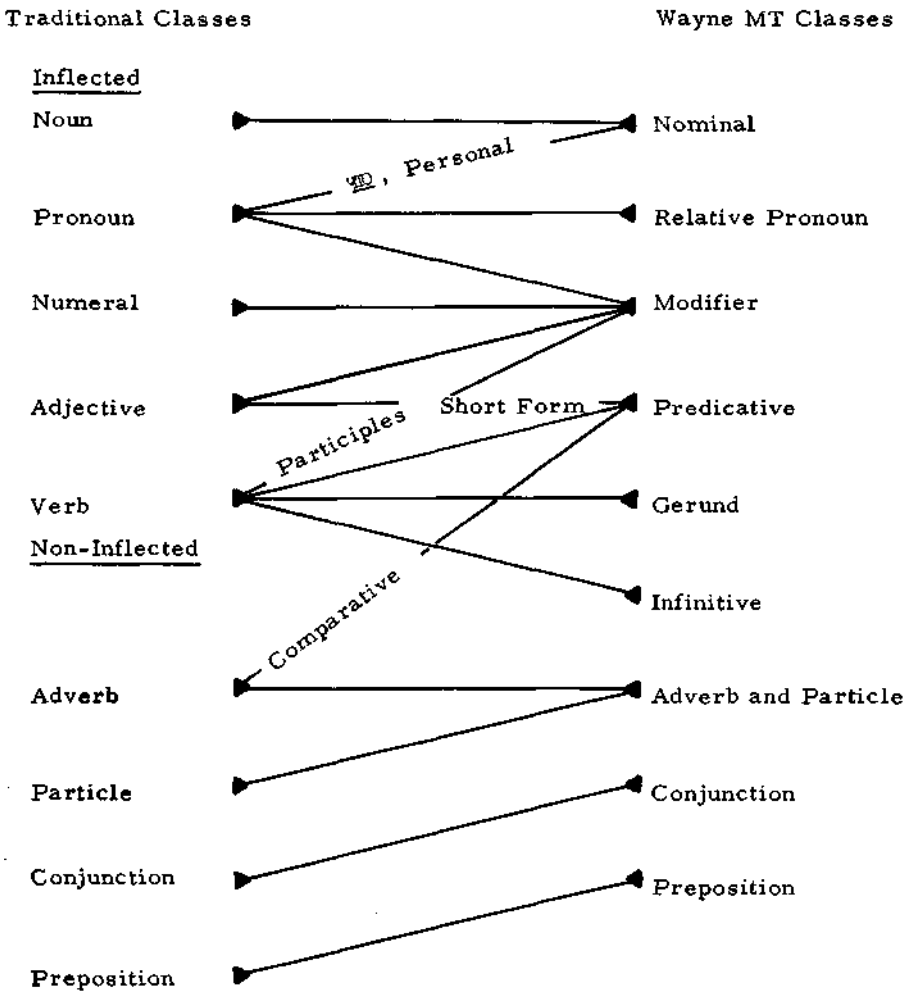


Figure 2 - WORD CLASS INFORMATION FOR SYNTACTIC CODING PURPOSES (Distinctions required in addition to agreement and government codes)
unmodifiable (pronouns)
ЧТО, ОН, ОНА
NOMINAL (nouns and pronouns)
modifiable (nouns)
not differentiated for nom/acc
РЕШЕНИЕ
differentiated for nom/acc
for both singular and plural
БРАТ
for singular only
ФУНКЦИЯ
differentiated for either number or gender
for number only (non-past of verbs)
ИМЕЕМ
non-third person
third person
ИМЕЕТ
for both number and gender (past tense, short forms)
can take БЫ (past tense)
БЫЛО
cannot take (others)
БЫ
ДЕЛАЕТ
PREDICATIVE (finite verbs, short forms, comparatives, dash)
undifferentiated
ВОЛЬШЕ
can be impersonal
БЫЛО
cannot be impersonal
ЧИТАЛ
governing participles, adjectives like
УДОВЛЕТВОРЯЮЩИЙ АНАЛОГИЧНЫЙ
MODIFIER
adjective, participle, certain pronouns
modifiable
ВАЖНЫЙ
unmodifiable
ВСЯКИЙ
non-governing
with normal agreement
НАЧАЛЬНЫЙ
with abnormal agreement (non-oblique numerals)
ПЯТЬ, ДВА
functionally ambiguous
И, ИЛИ, А, НО
punctuation
INFINITIVE
GERUND
CONJUNCTION
functionally unambiguous
finite
ПОТОМУ ЧТО
non-finite
ЕСЛИ, ЧТОБЫ
ADVERB AND PARTICLE
PREPOSITION
RELATIVE PRONOUN
КОТОРЫЙ, ЧТО
Adapted from Paul Garvin

Figure 3 - A COMPARISON OF GRAMMAR FORM CLASSES



In addition to a word-class code, agreement and government codes become part of the grammar code. The government code has two aspects: case government, and the less obvious prepositional government. The government codes relate to each other syntactic elements which are closely dependent rather than just loosely collocated. In regard to prepositional government, we have at present three criteria:

(1) semantic connection (how is the translation modified by the appearance of the preposition?)

(2) variety of prepositions that occur (the less the variety occurring with a given word, the greater the probability of dependent connection)

(3) degree of translation disturbance if the prepositional structure is omitted

Since we are going to use an IBM 709, all the information contained in the grammar code is coded in binary form. So far the entire grammar code occupies three 36-bit machine words. The actual binary code is such that each grammatical distinction is, insofar as possible, represented by a single binary bit. For this scheme we are indebted to the Ramo-Wooldridge Laboratories, who are working under an Air Force contract. An illustration of the details of how the individual words are coded is presented in Figure 4, while Figure 5 illustrates the coding of an actual Russian sentence taken from our mathematical text.

As an example of one of our program steps, we will now describe our nominal block routine, illustrated in Figure 6. When a nominal is detected in a left-to-right scan of the sentence, the nominal blocking routine is entered for the purpose of packaging the nominal with its preceding modifiers, including, possibly, adverbs modifying these preceding modifiers. The first question asks whether there is any string of modifiers preceding the nominal. Intermediate adverbs modifying members of the string are included in the package, since they do not terminate the search. Commas and ambiguous conjunctions (и, а, или, но, and others) are also skipped because their function is to separate the series of modifiers.

If there are no modifiers, then we conclude that the nominal block consists of just one element, namely, the nominal. If there

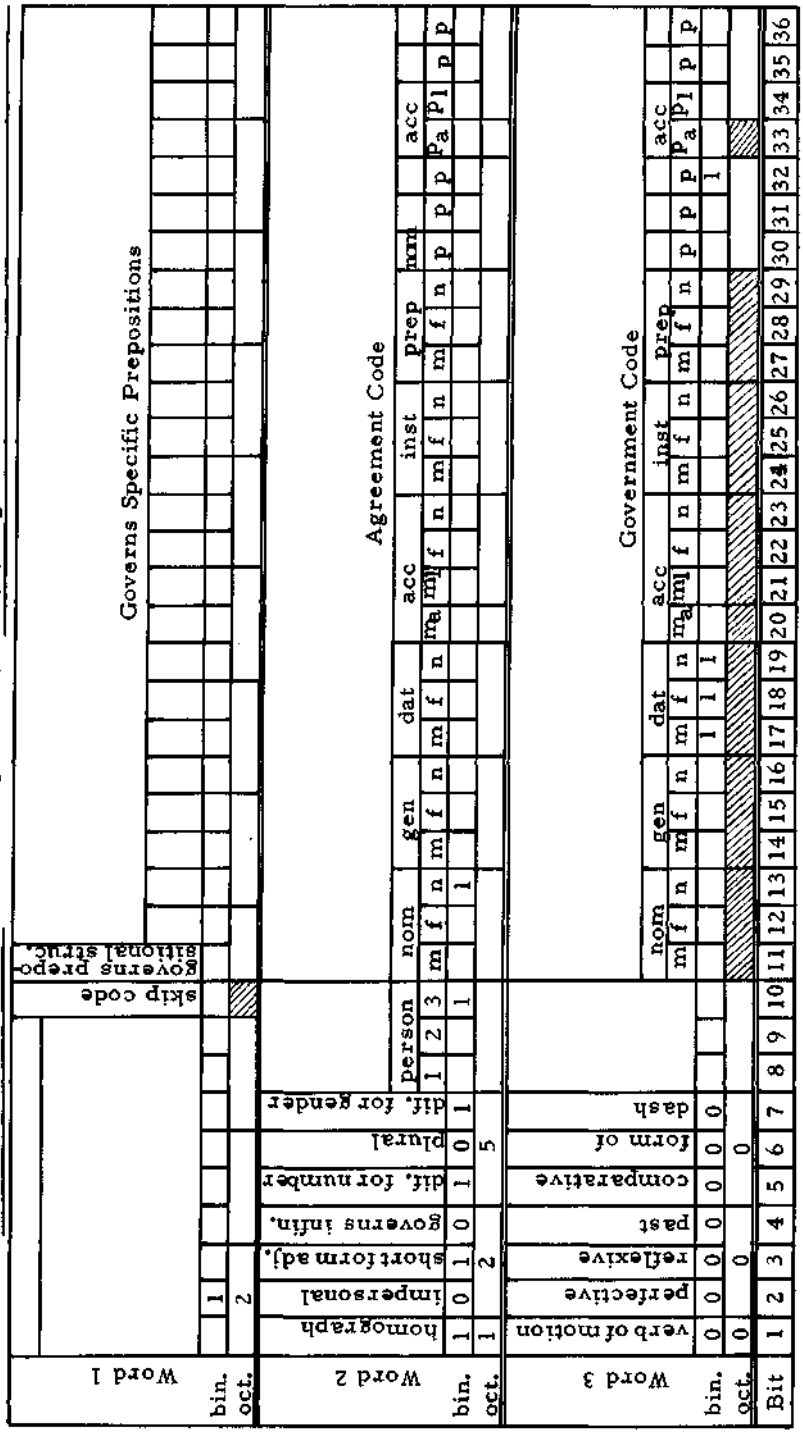


Figure 4 - DETAILS OF CODING OF INDIVIDUAL WORDS--PREDICATIVE

FIGURE 5

Illustration of Wayne MT Russian form Classes in a Sentence

A* V C C* C A C P M N
Хорошо известно, что, например, для гиперболических уравнений
N C* N N N N V N P M
существование и единственность решения задачи Коши имеют место без всяких
N N M N P N C C N
ограничений роста начальной функции на бесконечности, поскольку значения
N P N V A* P N M N P* M
решения в точке зависят лишь от значений начальной функции внутри соответствующего
N N C
конуса характеристик.

(It is well known that, for example, for hyperbolic equations the existence and uniqueness of the solution of the problem of Cauchy hold without any restrictions on the growth of the initial function at infinity, since the values of the solution at a point depend only on the values of the initial function inside the corresponding cone of characteristics.)

Explanation of abbreviations: N - nominal; m - modifier; R - relative pronoun; v - predicative; A - adverb, particle; P - preposition; C - conjunction, punctuation; * - homograph.

The following homographs are found in the above sentence:

хорошо -A or V

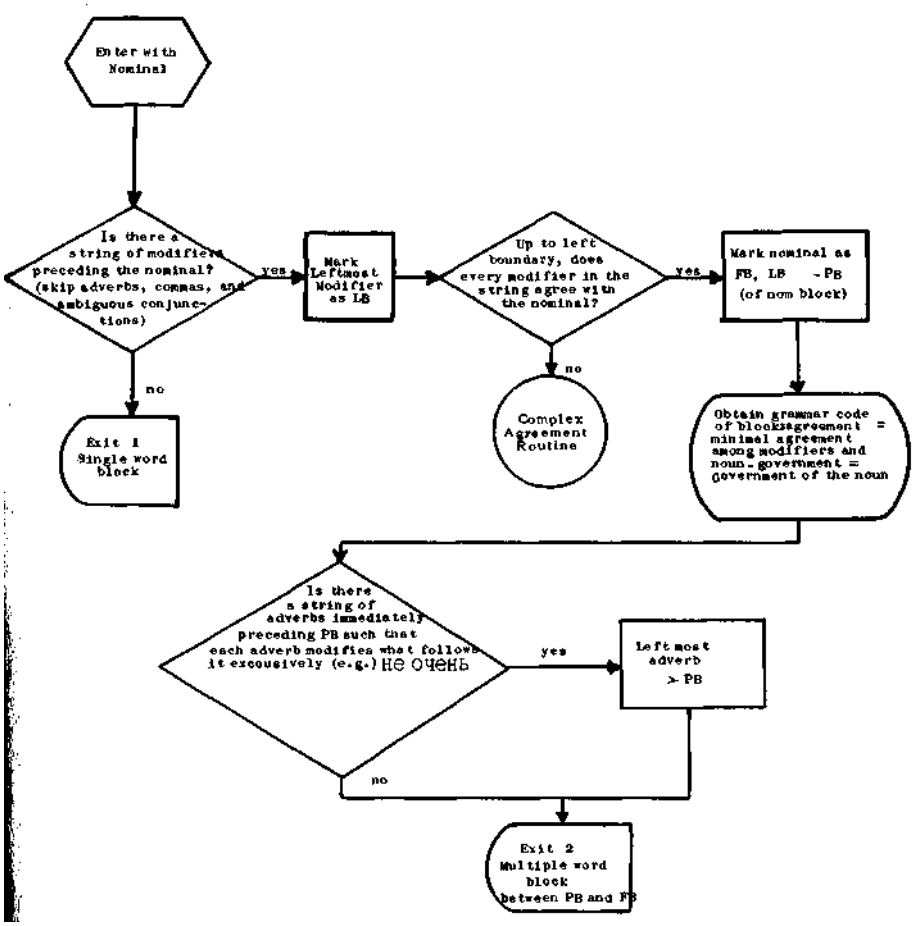
что -N, C, or R

и -A or C

лишь -A or C

внутри -A or C

Figure 6 - NOMINAL BLOCKING ROUTINE



LB: Left Boundary; FB: Final Boundary; PB: Preceding Boundary

Example: (для) очень гладких и бесконечно дифференцируемых функций
(for) very smooth and infinitely differentiable functions

are modifiers, then we must ask whether they agree with the nominal. If any of the modifiers fails to agree with the nominal (in number, case, and gender), then it is necessary to investigate this string in terms of more complex types of agreement; e.g., a numeral in the string of modifiers, or a compound nominal block, or adjectives whose nature requires them to be singular while modifying plural nouns, as in для одного или нескольких уравнений = "for one or several equations". These eventualities will be considered in the subsequent complex agreement routine.

If, however, the modifiers all agree with the nominal, we can mark the boundaries of the nominal block, with the leftmost modifier as the preceding boundary, and the nominal as the following boundary. At this point, the entire block is given a grammar code so that it may be treated as a unit. The grammar code is obtained as follows: the agreement code of the block is the minimal agreement among the modifiers and the nominal, and the government code of the block is the government code attributed to the nominal.

It may be possible to extend the left boundary of the nominal block to include preceding adverbs, if these adverbs are known to belong to the modifiers which follow them. If so, the preceding boundary is changed, and marked at the leftmost adverb satisfying the condition. In either case, we exit from the routine with a multiple word nominal block marked by preceding and following boundaries and grammar coded as indicated.

The nominal block routine just described and a number of similar routines, the purpose of which is to identify other properly labeled blocks comprising the sentence, are conceived of as preliminary passes to the main syntax passes. The purpose of the latter is to produce by the application of proper routines a sentence image, which in turn will yield, after being subjected to clean-up passes to resolve the few remaining lexical and morphological ambiguities, a better than word-for-word translation.