# SOME GENERAL QUESTIONS OF MACHINE TRANSLATION

I. K. BEL'SKAYA (Cand: Philol. Sci.)

*Machine Translation, pp. 3-32*

## 1. Linguistic prerequisites of machine translation

Already at the time of the first conference on machine translation (1952, U.S.A.), all the scientists interested in this problem had apparently concluded that such translation was possible. ,

The basic prerequisites for machine translation (MT) to be possible were then given as:

1. The language must be of the character of sighs enabling it to be compared with any other system of conventional signs (code, systems of mathematical symbols, &c.).

   Doctor Oettinger, in particular, considered that such a process as language translation was similar to processes like encoding telegrams or converting numbers from the decimal to the binary system, because "in each of these cases there is a recoding of a message by means of a new system of symbols " (1, p. 50).

2. The existence of a common logical basis for all languages which, it is supposed, makes it possible to discover a so far undiscovered " universal language" common to all languages, by abstracting from particular differences between languages (1, p. 22).

Professor Reifler, wrote "a relationship between different languages is only possible in cases where some common features are inherent in them. All languages have in fact a certain number of common features, Similar features are particularly numerous owing to the logical aspect" (2, p. 140). "In a language there are, of course, elements which are analogous (intuitive sense of style, emotional content, &c.)", wrote Professor Weaver—a reason for pessimism as regards the problem of literary translation. But as a written language is a system of a logical character this problem (the problem of MT--author) can be solved, at least formally (1, p. 27).

Both these prerequisites, it appears to us, give rise to serious objections.

First, translation from one language to another is not a special case of cryptography because in linguistic expression of thought the relationship of content and form is different from that in coding a text—considerably more complex, for here content and form are *mutually inter-dependent.*

A code* alters nothing in the content of the text to be copied, being a factor external and incidental to the text. As regards language, the position is different. As is well-known meaning is not only fixed ("coded") by means of the language but also formed. ,

> "Dictionary meanings of words are so bound up with the specific nature of the actual language in question that their universal, conceptual, logical content is overgrown on all sides with the characteristic forms and shades of meaning of the national genius of the people in question ".([2])

Thought based on *different* languages operates with *different* associations according to the laws of these languages. Consequently, there are grounds for asserting that a language is *internally* connected with thought, determining the " trains" of thought.

This makes a language a specific system different from any other system of signs and in particular, from code. This accounts for the "lack of correspondence" between language and code which have been noted even by the supporters of the above conception.

Thus A. G. Oettinger([1]) notes that "the symbols of a new code should so correspond with the symbols of the old code that the content itself remains unchanged". For codes this ideal is easily attained but it is completely unattainable

* If this word is used in its terminological meaning and not misused in the unjustifiable extended concept of the term “code” met with in some recent works.

in language translation, where "two codes are given and the problem consists in finding the relationship between them". Oettinger regrets that "the algorithm of translation from natural languages is unknown and a mutually non-ambiguous relationship is the exception rather than the rule ".

As regards the second prerequisite of MT, put forward by American investigators, let us note that it is obviously the result of *confusing* the categories of language with the categories of *logic.*

In attempting to separate the invariant components of a sentence in different languages, L. Wundheiler introduces the concept of "synonymous sentence" which he defines in such a way that the peculiar linguistic connexions and relations between the words in the sentence are eliminated.

"Sentences are synonymous" Wundheiler writes, if they contain identical information on certain relationships between the objects (designata) in question in the sentence considered. These relationships moreover are common components of synonymous sentences. Thus the two sentences "I gave a book to him" and "He was given a book by me" establish the same relationships between a certain person who gives the book, another person to whom the book is given and the book which is given.

We assume that in all languages there is an element denoting the action called in English "give" and that this element is used along with other elements denoting person giving, person to whom given and the thing given.([3])

It is quite evident that abstraction from all the linguistic peculiarities of all languages gives laws of logic* common to all thought, and translation from such a "general language" to any "special languages" requires that the latter be established in all their complexity and preliminary abstraction will not facilitate this task.

In the expression of thought in another language, L. Bloomfield writes, differences "will only concern structures (linguistic) of forms and their connexions and relationships *inter se* (their connotations)" (4, p. 278).

This, however, is precisely what language is—a determined structure of linguistic forms which are interconnected and this structure is intended to serve in the realisation of the human faculty of thinking and communicating thought. It therefore appears to us that the assertion that "the only basis permitting translation from one language to another" is the presence of common features in both([5]), is incorrect. Languages may be very similar in structure (any related languages) or very far apart (the Russian and Chinese languages); in the first case this facilitates the task of translation but in the second case the difference between the structures does not make the task impossible.

We consider that the possibility of machine translation depends on the following three factors: the systematic nature of the language, the objectivity of linguistic laws and the "formalness" of the language.

The recognition of the fact that a language, despite all its complexity, is not totally chaotic but has its own completed structural system, functioning according to its own definite laws, is the chief linguistic requirement for carrying out MT.

The laws governing the functioning of each language are objective and can be formulated as laws of grammar and laws of lexicology for the language in question; the laws governing the sounds of a language are reflected in the laws of phonetics.

In comparison with these concepts which are built up by means of words and based on them, all the categories of language including the dictionary meaning of words, which are closely connected with the grammatical meanings, are structural elements of the language and in this sense they are formal.

In a language there are no unformulated meanings and there is no formulation of what means nothing. The limitation of the meanings in a word and also any changes in its meaning (either dictionary or grammatical meaning) is formally fixed in the language by the different means of dictionary, grammar, and, partly, phonetics. Thus analysis of the formulation of a word (in the wide sense) gives all the necessary information for a correct understanding of the word in a sentence.

* Cf. Definition of logic: "logic is the theoretical science of the correct forms of thought" (Asmus, V.F. " Logic " Moscow 1947, p. 8).

Normally we have no doubt that a direct connexion exists between a given grammatical form of a word and the existence or characteristic absence from it of some material formulation. But the idea that the choice of meaning for a word with several meanings may also be formally determined appears less self-evident. Translators are frequently inclined to resort to "intuition", "the general content of the phrase" and such indefinite, non-formal reasons for selecting a particular meaning from several possible alternatives for a given word.

This characteristic error is due to the fact that the laws of a language as affecting the dictionary meaning are more complex than the grammatical laws. They are therefore much more difficult to systematise, it is harder to see the general behind the particular and formulate it as a law. Nevertheless the more exactly the lexicological laws are fixed, the more evident will become the connexion between the formal and ideal (cogitative) features of this connexion in the field of lexicology.

Work on carrying out MT on the linguistic level should assist greatly in improving the accuracy of existing definitions of fundamental linguistic categories since this work not only makes it possible but also imperative to make a careful linguistic analysis of a large amount of experimental linguistic material.

Attempts to produce an "analytical syntax system", an "operating syntax system", to carry out "structural analysis", which are characteristic of foreign linguistics, are due to a large extent to the fact that the contradiction and verbiage of the definitions of the fundamental linguistic categories make it impossible to establish uniform principles for the analysis of linguistic systems in a whole series of cases.

For the same reasons, Professor E. Reifler (1, p. 137, 5) regarded it as necessary to speak of "two linguistics", "traditional linguistics" and "MT linguistics", regarding the basic difference between them to be that the latter is very much more closely linked with practice and that its fundamental purpose is formal analysis for defining the meaning of words.

E. Reifler considers that the ideas of "traditional linguistics" are only acceptable for "MT linguistics" in so far as they can be verified as being suitable for it but he does not see that ultimately the ideas of the first and second linguistics, verified by means of the second, should coincide: from this point of view "impractical" ideas are simply wrong ideas, since they incorrectly or inaccurately describe the objective laws of the language.

## 2. Dictionary for machine translation*

In principle it is perfectly possible to construct a satisfactory dictionary for machine translation without contravening fundamental lexicological traditions established in relation to two-language dictionaries. In particular there is no great need to change the MT dictionary into a "dictionary of stems" (Cf. ([6])). A dictionary of words for machine translation also possesses definite advantages†.

Besides this the MT dictionary is a new type of two-language dictionary differing from the normal type both in structure and method.

The structural peculiarity of the MT dictionary lies in the separation within it of a certain number of structurally independent components.

First of all this means independent retention of the basic components of the two-language dictionary, the dictionary of the language translated and that of the language into which it is translated.

* We shall not dwell on questions of technically producing dictionaries for MT, since the subject of the present paper is the philological problems of automatic translation. For the other problems of machine translation we refer the reader to the papers: D.Yu. Panov "Automatic translation" M. 1958; 1. S. Mukhin "Trial automatic translation on BESM electronic computer" M. 1956; I. K. Bel'skaya, I. S. Mukhin "Automatic translation from English to Russian on BESM" (in the Press). (Material of conference "Ways of developing Soviet mathematical machine and instrument making ", March 1956).

† Thus the grammatical characteristic of words, given in dictionaries as initial information for subsequent grammatical analysis of the phrase (see below on "invariant characteristics of words") is possible for every word but is not always possible for stems.

Thus in our case the following dictionaries are recorded:

A. 1. English (maths. dictionary)
2. Chinese "(maths. dictionary)
3. German (maths. dictionary)
4. Japanese (maths. dictionary)

B. 5. Russian (maths. dictionary)

Within the dictionary for each language there is a division into two main sections:—

I.—Single-meanings dictionary
II.—Many-meanings dictionary

each in turn having two sub-sections:—

I*a*.—Dictionary of terms
I*b*.—Dictionary of single-meaning words in common list
IIa.—Dictionary of independent multi-meaning words
II*b*.— Dictionary of service words

This type of dictionary structure ensures its accurate and rapid working in MT, and also provides for the possibility of the non-uniform expansion of its different parts.

The main features as regards method of the MT dictionary are:—

First, in the MT dictionary each word is accompanied by a systematised description to allow of subsequent grammatical analysis of words in a phrase, i.e., gives the "*invariant characteristic of the word*".

Secondly, in the MT dictionary words obtain meanings which in many cases do not coincide with normal translations of these, but arise from subsequent comparison of two lexicological systems (language-source and language translation), *i.e.,* " *relative meanings of words"* are given.

Thirdly, cases where the word should not be translated into the other language as a separate lexicological unit are described ("fixed") as special meanings ("zero meanings") in the MT dictionary.

The "invariant characteristic of a word" also includes grammatical and semantic elements in the description of a word ("features") which, by determining the type of word-variation and the character of word-combination, retain a constant meaning in any use of the word. In each language grammatical categories and semantic features of this type form a certain definite system of characteristics, constant for the language in question, and this can be given in the MT dictionary along with the translation of the dictionary meaning of the word. *The systems of invariant (or dictionary) features* are contrasted with the systems of variant (contextual) features of a word the meanings of which vary according to the function of the word in the sentence and consequently only determined by analysing the sentence.

Invariant features of nouns in Russian are the type of declension, type of root, categories of grammatical gender, animation, appellativeness, pronoun-ness and also special features in formation and membership of some separate lexicological group, while the variant features are number and case only.

The invariant features of a word form the basis for grammatical analysis of the word in the sentence and the variant features its purpose.

The view that it is necessary to show some grammatical information on words in a dictionary has been expressed both by compilers of ordinary dictionaries (7, p. 102) and also by workers studying the problem of machine translation([8]), (1, p. 47,61). We have attempted to systematise dictionary grammatical information on words since this is of fundamental importance in machine translation.

As regards the character of the meanings which the word is given in the MT dictionary, their specific nature is that they fix, with greater consistency than in ordinary dictionaries, the translation of the lexicological system of one language into that of the other language.

The large number of misunderstandings and distortions in translation, against which translators are usually warned (see in particular([9]), ([10]), ([11])). is accounted for by the "unreliability" of two-language dictionaries and their unsuitability for automatic ("unthinking") use.

In translating with an ordinary dictionary, a translator practically never confines himself to one meaning from the choice given by the dictionary (we are speaking of terms) but is forced to "think out" the meaning which he actually writes in the text of the translation. In this case what happens is nothing else but the replacement of some "prompting" meaning from the dictionary by the meaning which has been fixed in linguistic practice as the constant translation of the meaning in question (and sometimes the use) of a foreign word.

In our opinion, it is just these meanings which are actually used and which a word takes on in translation into another language, which should be fixed in any translating (two-language) dictionary. We call them *relative meanings of words* in distinction to *proper* meanings which words possess within their own language and which are fixed by norms and encyclopaedic dictionaries of these languages.

The defect of existing two-language dictionaries from the point of view of using them in translating practice in general and in MT in particular is their inconsistency in giving the relative values of words: in many cases translation of the words of another language is substituted by a translation of the contents of the encyclopaedic dictionary of this language.

Whereas in the normal two-language dictionary, the absence of relative meanings can be compensated for by the "intelligence" of the translator, with an MT dictionary working without human intervention, this disadvantage may reduce to almost failure the working of the dictionary. Therefore, the necessary "thinking" should be done beforehand and fixed in the dictionary in the form of relative meanings of words.

A frequently occurring case of a relative meaning is the "null meaning" of a word. In this case the word without undergoing any change at all in its structural significance (in the sentence to be translated) loses its lexicological independence and has no independent lexicological equivalent in the translated sentence.

The most important (actual) are "null meanings" for subordinate words (in the meaning of which the amount of lexicological meaning is generally not great), but a word having a full meaning may also have a "null meaning" in particular those words which are components of idiomatic expressions.

The question of translation of idioms into another language which has frequently been considered by investigators as giving rise to the greatest doubt (1, pp. 183-194)([12]), in our opinion can be satisfactorily solved by including the idiom in the dictionary of words with multiple meanings*. In this case an indication of the idiomatic use of the word is included in the general system of indications ensuring correct translation of words with multiple meanings.

In conclusion, let us note that the problem of the MT dictionary, with which the consideration the possibility of automatic translation has sometimes begun, is not the most difficult.

Whereas in the early stages, the greatest doubts were expressed as to the possibility of "remembering" a dictionary of sufficient size (astronomical figures —up to 30 million words were mentioned in this connexion([12])), latterly the problem of multi-meanings of words has been regarded as a similarly insoluble problem.

In our opinion both the above problems—the size of the dictionary and multi-meaning of words—can be solved satisfactorily for all languages by a combination of two methods:

*(a)* Division of the MT dictionary into a series of "special dictionaries" corresponding to different spheres of human activity, in our case, corresponding to the different branches of science;

*(b)* Use of contextual (functional-semantic) analysis of words with multiple meanings.

It can be asserted that in this way we should attain not only reduction of the actual† volume to 4-5 thousand words but also restrict the rules regulating words with multiple meanings.

* Section II *a* of the MT dictionary.

† The real volume of the dictionary is the volume for which the dictionary can ensure independently the translation of any text in the given field.

## 3. Questions of grammatical analysis

Recently questions-of grammatical analysis in MT have attracted the attention of a large number of research workers in different fields. It should be stated that our point of view in this matter differs from the majority of alternatives at present being suggested for the solution of the problem in at least two respects.

Firstly, we do not propose a break with traditional linguistics on the fundamental positions and the solution of the problem of structural analysis of phrases is viewed by us rather as a matter of obtaining concrete results from observations of a linguistic character than as one of using mathematical interpretation of linguistic phenomena.

Secondly, we regard our work in carrying out MT not so much as some kind of conventional-symbolic activity (which could later be used for more general purposes than translation) but rather as an attempt at complete formalisation of actual translation in all its concreteness.

For this purpose we set up experimental schemes for grammatical analysis of phrases for MT into Russian from English, Chinese, Japanese, Russian and German and the description of these forms the main part of the present collection of papers.

We deal below with the principles of grammatical analysis of a sentence which are applied in analytical MT systems from the languages named.

For a grammatical analysis of five-language systems, English, German, Russian, Chinese and Japanese, it appeared possible to use a similar system of division of words into nine lexico-grammatical classes

1. Verbs
2. Nouns
3. Numerals
4. Adjectives
5. Adverbs
6. Prepositions*
7. Conjunctions
8. Particles
9, Introductory words

The principle of the division of words into these classes coincides with that underlying the division of words into parts of speech. There was therefore no need to abolish the traditional names of the parts of speech and replace them by structuralistic signs. It was only to make the categories of the parts of speech more precise.

The difference between parts and particles of speech is reflected in the classification used by giving the latter the specific syntactic feature of "minimum structural significance".

The elimination of pronouns as an independent part of speech is not contrary to definite linguistic tradition, in particular English grammatical tradition (but), is in conformity with one of the most consistent conceptions of the parts of speech in modern linguistics.(13) The possibility of using it in MT evidently confirms the correctness of this conception.

We distinguish the category of pronouns for such parts of speech as nouns, numerals, adverbs and adjectives.

The class of adjectives is also different in composition from the traditional, owing to the inclusion in it of a category of words usually called "ordinal numbers". The reason for the change was that the "ordinal numbers" lacked any grounds for being regarded as numerals, other than semantic grounds.

Also for structural-semantic reasons, "interjections" were combined in a single class with introductory words.

The class of adverbs contains only proper adverbs, *i.e.,* adverbs of place, time and iteration. Words traditionally called "adverbs of manner of action" including "qualitative adverbs" are classified as adjectives with an adverbial function.

In a number of languages, adjectives in the adverbial function can be specifically formulated:

*e.g.,* in Russian настоящий — по-настоящему (actual, actually)
механический — механически (mechanical,
mechanically), &c.

In English the suffix -ly is used for this purpose: mechanically, grammatically.

In German—the absence of case endings of the adjective.

* The postpositions in Chinese and Japanese can be classed as postpositive prepositions owing to the similarity of function to prepositions.

However, even in these languages the special adverbial form does not necessarily accompany the adverbial function of adjectives. Thus, "qualitative adverbs" in Russian are not differentiated morphologically from the short form of the corresponding adjectives. English abounds in "homonymic" adjectives and adverbs of the type "hard", "easy", &c.

*e.g.*, hard work, to work hard, hard to choose
easy task, easy to understand, take it easy, &c.

Adjectives are used in the adverbial function in Chinese with complete freedom.

Comparative analysis of the five languages, Russian, English, German, Chinese and Japanese showed that the degree of structural separation of adjectives in the adverbial function in these languages is insufficient and so different that relation of the languages in that plane is devoid of any regularity. A regular relationship is only possible for certain very general categories and the category of adjectives is not such a category.

The grammatical categories within each part of speech are fixed in the form of variant (context) and invariant (lexicological) features of words.

Amongst invariant features, those shared by the five languages (universal features) are distinguished from those which are specific to the individual languages (specific features). ,

The following invariant signs are universal:

| | |
|---|---|
| For verbs | 1. Modality; 2. Transitivity; 3. Conjunctivityl |
| For substantives | 1. Pronoun-ness; 2. Animation; 3. Appellative-ness; 4. Collectivity; 5. Conjunctivity; 6. Verbal character. |
| Numerals | 1. Pronoun-ness. |
| Adjectives | 1. Pronoun-ness; 2. Conjunctivity; 3. Quality; 4. Semantic shading of ordinal adjective. |
| Adverbs | 1. Pronoun-ness; 2. Modality; 3. Negation; 4. Semantic shades: adverbs of time, place and iteration. |
| Conjunctions | 1. Type of connexion: uniform or -non-uniform; 2. Semantic shading: condition, purpose, time. |
| Particles | 1. Subordination; 2. Semantic shades: negation, intensification. |

Of the invariant features enumerated only the following need special explanation:

The Conjunctivity of verbs, the ability of a definite group of verbs to occur regularly in the role of connecting verbs; in Russian this group includes the verbs быть, являться, называться, казаться, становиться, оказываться; in English—become, turn, grow and also seem, look, feel, smell, &c. In Japanese ある (aru) apy, なる (naru) нару with its stylistic equivalents and verbs derived from it; in Chinese 是，为，乃，名为 &c.

Conjunctivity of nouns is a constant characteristic of a definite group of pronounial substantives appearing as conjunctions (Russian: кто, что, &c.); the necessity of separating this category of substantives is due to their specific syntactical functioning as conjunctions: they always begin a sentence*.

Modal verbs in Japanese are not only the verb (dekiru) "to be able" but also those "restored" from verb-suffixes of a verb corresponding to modal verbs in the other languages.

Example 讀 める (emeru) "I can read" where 讀. is translated by the verb in the infinitive form "to read" and from the ending める the modal verb "I can" is "restored".

Modity of adverbs in the generalisation of special features in semantics and use of a group of adverbs often called "categories of state". In Russian: "можно, нужно, надо, нельзя" (can, must, needs to, must not†..

* The text included between full stops, question or exclamation marks are termed *phrases*. We term *sentence* simple sentences, *i.e.,* those having not more than one non-uniform predicate.

† The group of words ending in -o of the type хорошо, плохо, далеко (good bad, for), &c. which are classed in this category of state in the Academic Grammar of Russian are not included in the category of modal adverbs (in Russian) because of the retention of a distinct formal relation with the basic form of the adjective: хорошо-хороший, плохо-плохой, &c.

In Chinese:

| | | |
|---|---|---|
| 難 - difficult | 難以-difficult | 容易-easy |
| 可以 -possible | 不可以 - impossible | 不能 - possible |
| 不可 -impossible | 可能够- possible | |
| 不能 - impossible | 不難- not difficult 可 | |

In Japanese, this category includes words normally regarded as adjectives in the finite form which are joined to existing forms of verbs, *e.g.*, 讀みにくい (yeminikuy) is translated by two words " to read" and "difficult", the second word being classified as a modal adverb.

Finally, as regards the division of conjunctions into uniform and non-uniform it is necessary to explain

> *Uniform conjunctions* are those joining uniform "terms of the sentence".
>
> *Non-uniform conjunctions* are those which join the whole sentence independently of the type of connexion of this sentence with the remaining sentence of the phrase. Non-uniform conjunctions are the marks by means of which the phrase is divided into sentences. This division of the phrase is so important and difficult a problem in MT that with conjunctions carrying put two functions simultaneously, the union of a uniform sentence, it is more important to stress the function of the non-uniform, conjunction.

For example in the phrase "Let us assume that *f (x, y)* is less than *M* and *f(x,y)* satisfies the condition . . . . " (M46), the conjunction "and" is classified as non-uniform, joining the sentence "*f(x, y)* satisfies the condition . . . . "

The conjunction "and" has an analogous characteristic in the following case:

> " From these results we see that the initial error is very small, that in increasing *n* the error increases slowly *and* that finally . . . . it begins to change Sign at each step " (M64).

A number of conjunctions, normally called compositive, can be used equally in both functions. In the dictionary these conjunctions are regarded as words with multiple meanings, distinguished by their grammatical characteristics.

Thus in the sentence:

> "The whole meadow *and* the shrubs round the river were submerged in the spring floods *and* between Zhukov and that side the whole space was completely occupied by a huge flood " (Chekhov, vol. 8, p. 231).

The conjunction in the first and third cases are called “uniform" and the second "non-uniform".

As regards invariant features which are special to particular languages we shall merely explain that indications of the membership of words of different classes in a particular structural-semantic or morphological group given to the word in the dictionary, fall into the same category.

Among the variant (contextual) grammatical features we distinguish between "auxiliary" and "final" (features). The latter are none other than Russian variant features*: analysis of a substantive of any language, whether Chinese, English or Japanese, is directed towards the elucidation of the case and number of the Russian equivalent substantive, and analysis of an adjective similarly for elucidation of the gender, number and case and other variant features of the Russian adjective.

In the system of MT analysis differences are observed in languages experiencing the necessity of fixing the intermediate results of analysis in the form of "auxiliary" variant features and those not experiencing such necessity.

Chinese and German should be placed in the first group, and English Japanese (partly) and Russian in the second,

* Translation from Russian does not constitute an exception for in this case Russian-Russian analysis occurs—due to the special position of the Russian language in our system of MT in which Russian is assumed to be used as the intermediary language.

The auxiliary features in Chinese are the syntactic features which are obtained in the first stage of the analysis. In translating from English, Japanese and Russian only the final features are fixed.

## 4. General characteristics of schemes for translation

The process of machine translation of a phrase falls into two main parts:

*A.*—Finding the words in the MT dictionary by the method of applying in turn all the words of the phrase to the dictionary,

*B.*—Treating the phrase with translation schemes, These latter fall into three cycles as follows:*

Cycle I.—Dictionary schemes:

1. Division of phrase into words
2. Finding dictionary form of words
3. Finding part of speech of "unknown word"
4. Syntactic analysis of "formulae"
5. Distinguishing homonyms
6. Analysis of words with several meanings
7. Application to Russian synthesis

Cycle II.—Analysis schemes:

1. Functional analysis of punctuation signs
2. Compacting phrases into sentences and clarifying distinctions in sentences
3. Syntactic analysis of sentences
4. "Verb" scheme
5. " Numeral " scheme
6. " Substantive " scheme
7. " Adjective " scheme
8. Change of word order in translated phrase

Cycle III:

1. Word-forming scheme
2. "Verb" scheme
3. "Adjective" scheme
4. "Substantive and numeral" scheme
5. Stylistic editing of translated phrase

The schemes in Cycle I either operate on the MT dictionary (Schemes 1-4) or extend it (Schemes 5-8).

In different languages the schemes for this cycle are different. Thus, Scheme 1 is only used in Chinese and Japanese in which the text of a phrase, which is written without breaks, is divided into the minimum lexicological units—words.

Scheme 2 of the same cycle is important for all the languages in question, except Chinese where the dictionary and contextual forms of the word coincide. In the other languages the contextual word form is replaced by the dictionary form of the word if they are found to differ, through application of the word to the dictionary proving to be without result.

Schemes 5 and 6 of Cycle I in the Russian schemes for translation are replaced by the scheme "Analysis of inversions ". Due to the specific nature of the Russian dictionary which has no section dealing with "words with many meanings",† translation of idiomatic word-combinations in Russian is secured by means of the scheme "Analysis of inversions".

For the remaining languages, the following main cases of lexicological-grammatical homonymy‡ are brought under Scheme 5.

(1) Verb-noun, (2) verb-adjective, (3) noun-adjective§, (4) adjective-adverb, (5) preposition verb. More complex cases of homonymy (and less

* In enumerating the schemes, the order in which they are used for work in machine translation is retained. Certain special features of the languages in the sequence of schemes for Cycle II are stipulated below.

† See paper by Nikolayeva "Analysis of Russian propositions".

‡ It should be stipulated that "for technical reasons" only homographs are termed homonyms, all other cases of homonymy being excluded.

§ In this case in Chinese two alternatives arc analysed separately: conversion of qualitative adjectives and conversion of relative adjectives.

regular ones) are considered individually in the system of analysis of words with many meanings. It should be noted that the importance of the types of homonymy enumerated are different for different languages: thus type (5) (preposition-verb) is of importance in Chinese but not in English or German.

Scheme 6 of Cycle I, is the "working part" of the dictionary section for words with several meanings. The result is to indicate the number of the equivalent for words with several meanings in the Russian dictionary and also to define the invariant grammatical features of the word in question more exactly.

Schemes 3 and 4, operate the dictionary for words and conventional notations which are not included in the MT dictionary. These may be words which rarely occur in texts in the branch of science considered or are not at all characteristic, such as words from other fields. In the first stages such words may not be encountered by the compiler of the MT dictionary and so remain for some time " unknown words" for the machine.

The presence of "unknown words" in a sentence makes grammatical analysis of the sentence very difficult and in a number of cases, where there are a large number of such "blank spaces", it may be impossible. In the system of dictionary schemes, therefore, Scheme 3 is included which enables us to determine the part of speech of a word, the translation for which remains unknown. The scheme depends on analysis of the morphological form of the "unknown word" and its syntactic connexions in the phrase in question.

For a scientific-technical text, mathematical texts particularly, the existence of conditions of notation-formulae, mathematical signs, &c., are quite important and these, not being words, cannot be put in the dictionary. In our terminology we name these notations conventionally "formulae". The fact that the syntactic functions of formulae in a sentence are heterogeneous and far from equivalent from the point of view of grammatical analysis of the phrase to be translated gives rise to the need for a special scheme, clarifying the syntactic function of each formula in the phrase.

There is a certain difference as between languages in the volume of work involved in this scheme. In Japanese the "formulae" scheme includes syntactic analysis of Arabic numbers and also proper names which are not written in hieroglyphs or kana (the Japanese alphabet) since in regard to a sentence in Japanese they are "foreign words" no less than the conventional notations which we call "formulae".

The last dictionary Scheme (7), depends to a great extent in its operation on the results of certain of the schemes in Cycle II. It therefore begins to operate later: after all the schemes in Cycle II directly before the schemes in Cycle III.

The scheme consists of four separate parts corresponding to the four languages:

(1) English appendix (application)
(2) Chinese appendix
(3) Japanese appendix
(4) German appendix*

The object of the scheme is to prevent incorrect formations in Russian particularly "incorrect" formation of verbal substantives, and also the formation of both participles in the case of breakdown of the correlation of the corresponding participial forms of the two languages dealt with.

Analysis schemes (Cycle II) include four main Analysis schemes and a number of auxiliary schemes.

'The main Analysis schemes are Schemes 4, 5, 6 and 7 which deal successively with the first four classes of words: verbs, numerals, nouns and adjectives. These schemes are directly related to the synthesis schemes. The group of additional schemes comprises Schemes 1, 2, 3 and 7 which carry out the supplementary operations of Analysis, which prepare or formulate the results of working of the fundamental schemes.

Scheme 1 of Cycle II carries out analysis of the syntactic functions of all punctuation marks in the phrase to be translated apart from "finish signs", *i.e.,* full stops terminating it and also interrogation and exclamation marks.

* At present only the first part of this scheme has been completed.

The task of functional analysis of punctuation marks is simplified by linking machine translation to the framework of scientific-technical texts, in which the use of some punctuation marks is limited (hyphen) or altogether excluded (comma-hyphen),

The main objects of Scheme I analysis are commas, the most frequently used and most polyfunctional punctuation mark in all the languages studied. The remaining punctuation marks frequently analysed by Scheme 1 are conventionally (for ease in programming) regarded as variations of the comma. "Comma" with the feature "hyphen", comma with the feature "drop like" (in Chinese), &c., are differentiated.

The main functions of the comma are three in number*:

1. Division of a sentence into clauses
2. Introduction of a uniform term into a clause
3. Separation of words or groups of words from a clause†

The similarity of the first two functions of the "comma" with the functions of conjunctions makes a substantial saying in programming translation schemes by adding "comma" to the conjunction section. This curtails considerably the number of working checks in the Analysis schemes and enables the results of functional analysis of "commas" in the phrase to be recorded in a more uniform manner, without resorting to the introduction of new features.

Scheme 2 of Cycle II establishes accurately whether the sentence,‡ introduced into the machine for translation, consists of one clause or of several. In addition, in the same scheme the limits of separation of groups of words within the clause, are established. Thus the structural units within the limits of which the whole subsequent grammatical analysis will be carried out, are determined.

Such units are clauses (the fundamental unit of analysis) and so-called independent separations by which we understand any separations of words or groups of words in a sentence (apposition, participle and verbal participle inversions, introductory phrase) excepting introductory words§.

In a sentence consisting of several clauses with independent separations within them, further analysis *(i.e.,* after Scheme 2) is carried put in such a way that initially all the remaining schemes of analysis work in the limits of the first clause||, temporarily freed from "independent separations" if there are any such in it. The same Analysis schemes then work out these separations. Only after this does the analogous working out of the second clause begin, then the third and so on to the end of the phrase. The clauses are treated independently of each other except in cases where there are words of the nature of pronouns, their meaning being made clear by comparison with the previous clauses. Another case of going beyond the limits of the clause is analysis of the form of a verb-predicate in a proposition where the rule of " tense agreement" may apply.

Experience has shown that in all languages the number of clauses in a sentence may be determined in the same manner; by the number of non-uniform predicates. But the fixing of the limits of the clause requires individualisation by languages. The laws governing the separation of words within a clause are also individualised according to the language.

In this connexion, the position of Scheme 2 in the series of other analysis schemes is different in different languages and is determined chiefly by the degree of morphological formation (formedness) of the verb. Japanese and Russian verbs are morphologically formed so completely that the scheme "constitution of the sentence from clauses" can work as one of the first in the system of analysis.

In English and especially in Chinese it is in many cases difficult to find the predicate in a sentence without previous syntactical analysis of the clause (in

* It should be noted that in many cases a combination of the,functions given is. characteristic of the "comma" in a phrase. Cf., *e.g.,* the analysis of punctuation marks in Russian (see paper by T. M. Nikolayeva at the conference on MT, Moscow, 1958).

† The character of the separation may differ in different languages. Thus in Japanese the class of "separable commas" may include the "thematic comma", peculiar to Japanese. For greater detail on the "thematic comma" see paper by M. B. Yefimov at the conference on MT, Moscow, 1958.

‡ Here, as always, we use these terms in the sense already indicated above.

§ The exclusion of introductory words from the class of "independent separations" is due to the fact that introductory words being invariable, do not require any analysis.

|| Here it is kept in view that analysis of the sentence is carried out in the direction left to right.

Chinese) or general analysis of the verb (in English). Thus in Chinese, Analysis of the scheme "constitution of the sentence from clauses" is preceded by a special combined scheme "syntactical analysis of the clause" (Scheme 3)[*] and in English Analysis the main scheme of analysis is Analysis of the "verb".

A special feature of Russian Analysis is the combination of Schemes 1 and 2 of the Analysis owing to the fact that in Russian, separation of phrases into structurally independent parts depends in all cases on the use of certain punctuation marks.

The main Analysis schemes "Verb", "Noun"; and "Adjective" differ substantially between languages as regards the content of the analysis carried out in them[†]. Common features in these are similar methods (directedness) in working out final variant features, which are transferred directly to the schemes of Russian Synthesis and also common methodological principles of Analysis.

A dominant feature in all languages[‡] is syntactic analysis of words since, for translation it is necessary to establish something in the nature of "relative" grammatical characteristics from "particular" grammatical characteristics. We do not propose "relative" and "particular" grammatical characteristics because in this case the position is different from that in the case of dictionary meanings and the similarity in terminology might lead to erroneous conclusions.

Grammatical analysis makes clear the grammatical (variant) features of a Russian word in a Russian sentence, grammatical analysis being carried out on the corresponding word or words in the clause to be translated. In the case of both words we are speaking of "special" features, *i.e.,* those which the word and its equivalent have in the respective languages.

The "unambiguity" of syntactical analysis is ensured by means of morphological and partly by semantic analysis.

Attempts to exclude semantics and morphology from analysis of language structures substantially increases the number of "insoluble" (synthetically) problems.

The most usual case in which the insufficiency of a purely syntactic approach to analysis is clearly apparent is the precise statement of the determinant where there is participial inversion or a subordinate determinant clause and several nouns corresponding in position to the determinand.

1. "Аналогично можно интерполировать по значениям функций . . . , помещенным в одной какой-нибудь колонке таблицы" (Крылов, "Численные методы математического анализа).

Example I. "In an analogous manner it is possible to interpolate in respect to meanings, located in any column of the table" (Krylov "Numerical methods of mathematical analysis".)

In this example, the matching of case endings of the words "meanings" and "located" make it possible to determine the word defined by the participial form (inversion) while the purely syntactic norms of Russian permit relation of the subordinate determinant clause (and equally the participial inversion) to either of the two nouns preceding it.

Example 2. ". . . . using values of the function $f(x,y)$ located in the same line as $x=a_v$, we are able by the normal methods of interpolation worked out for functions of a single variable to calculate .the values of $f(a_v, y)$ *(ibid.).*

Example 3. "For equations in partial derivatives the number of magnitudes which have to be stored in the memory rapidly increases as the network diminishes *(ibid).*

Similar difficulties are encountered in the other languages.

Example 4. Another method for determination of the real roots of a transcendental equation which does not converge in all cases, consists in expanding the functions (S. & B., p. 15).

—(Другой метод для определения вещественных корней трансцендентного уравнения, который сходится не во всех случаях, состоит в разложении функций . . . )

[*] Scheme 3 "Syntactical analysis of the clause" is only important in MT for two languages, Chinese and Russian and its purpose in these languages is different, (See paper by V. A. Voronin at the MT conference, Moscow 1958, and in the paper by T. M. Nikolayeva mentioned above.

[†] See papers by V. A. Voronin, M. B. Yefimov and others at the conference on MT, Moscow, 1958.

[‡] The position is somewhat different in Russian Analysis due to the special position of Russian in our MT system, owing to the highly neutral nature of the results of analysis in relation to other languages.

Example 5. Gauss' scheme is a perfectly general systematic procedure for the *elimination* of the unknowns particularly adapted to slide rule: use and easily remembered (S. & B., p. 16).

(Схема Гаусса является достаточно общим систематическим приемом для исключения неизвестных, особенно пригодным при использовании счетной линейки и легко запоминаемым).

Solution of the problem is accomplished by resorting to analysis of the morphological form of the words (Examples 1--4) and where this does not help (Example 5) or does not solve the problem definitely (Example 4), to semantic analysis.

In MT semantic analysis presupposes preliminary structural-semantic classification of words as a result of which the semantic connexion of the word is fixed by its relation to a particular lexicological-semantic group. In schemes of grammatical analysis, me presence or absence of structure-influencing features in the semantics of a word are established by simply turning to the number of the structural-semantic group of the word, *i.e.,* not more complex than finding out the presence or absence of any grammatical inflexion in the word.

Special mention should be made of the question of the method of analysis of the context of the word in the phrase.

In many cases, precise grammatical description of the word makes it necessary to go outside the word in question (to analyse neighbouring words) preceding or following it. But this kind of check practically never involves the word literally preceding or following and as a general rule it is necessary to "slip in" a number of syntactically dependent words on the machine. "Passing through rules" which apply in the cases are based on the possibility of classifying all words in a sentence in the following three categories:

(1) Words of minimum structural significance—

Adverbs, particles, introductory words (not separated by punctuation from the sentence and also formants (in Chinese); "uniform groups in a sentence".*

(2) Words of second grade structural significance—

Any word or group of words (not included in Category 1) occupying a position of determination in the sentence to any word in the sentence.

This category of words is most varied and definite differences are noted in different languages.

(3) Words of first-grade structural significance—

These remain after the exclusion of words of the first and second categories.

In searching for words in the first category it is not allowed to pass the word through a single one of the categories mentioned. In searching for words in the second category only words of the first category are passed. In searching for words of the third category, maximum use is made of the "pass rule". Words are passed both in the first and second categories.

An entirely specific category consists of words of "zero structural significance". These are words not existing in the phrase to be translated but occurring in translation in the case where a group of words in the translation corresponds to a single word in the original. All words except the guide-word in such a combination of words has the feature "absolute passage" which means neglecting these words throughout the analysis of the words of the phrase in the original. The concluding phase in the operation of each of the four basic schemes of analysis is checking for the presence, in the translated sentence, of words corresponding with parts of speech with the "absolute passage" feature. Where the result is positive, the missing grammatical features are worked out for the words found.

Finally the last Analysis scheme, the "word order changes" scheme, is for the purpose of checking whether the order of the words in the translated (Russian) sentence corresponds to the norms for Russian.

It should be said that the problem of word order is of different importance in the different languages.

* We term "uniform groups" groups of words introduced into a sentence by a uniform conjunction or uniform comma. The limits of such groups are the uniform conjunction or comma on the one hand and the basic word which is introduced, on the other.

Changes in word order in translating from English to Russian are partial rather than fundamental in nature. Matters are somewhat more complicated with German and particularly with Japanese. The sequence of syntactical groups in the last mentioned language is substantially different from that in Russian and this necessitates a special group of analysing schemes combined in the "Scheme for word order changes".

The word order is of primary importance in translating into such languages as Chinese*.

The Synthesis schemes (Cycle III) include the basic rules for word variation in Russian and some rules for forming words. The schemes do not work out all the words in a Russian phrase but only those which are variable in Russian, *i.e.,* verbs, nouns, numerals and adjectives.†

The four basic schemes of Synthesis treat the above classes of words in succession. Scheme 1 ("word forming") is of an auxiliary nature completing dictionary operations with words.

In Scheme 5 of the Synthesis, work on the compilation of which is at present proceeding, correlates observations relating to the fundamental rules and methods of stylistic editing of the scientific technical text.

In the basic schemes of Synthesis (Schemes 2, 3 and 4) variant grammatical features which words receive in the Analysis schemes are realised. For this purpose, the dictionary ending of the word which does not correspond to the requirements of the variant grammatical features, is replaced by other endings satisfying these requirements. In a number of cases, e.g., in forming the participial forms of the verb the new ending can have a suffix added.

Formation of such grammatical categories as "verbal substantive" and the four participial forms of the verb is carried out in two stages: in the scheme "verb" the dictionary form of the verbal noun or participle is formed from the infinitive of the corresponding verb; subsequently in the schemes "noun" and "adjective" the word form so obtained is transformed along with other nouns and adjectives.

In principle the Synthesis schemes are so constructed that they embrace all types of CORRECT forms within the class of any part of speech. "Exceptions from the rules" including more or less individual order of roots, peculiar case endings of individual words or groups of words, &c., are present in the schemes only in the form of queries (checks) relative to the membership of a word in any of the special morphological groups. These groups combine words of the same part of speech diverging from the "regular" type of formation in a particular grammatical category. For words of this kind the Synthesis schemes provide individual rules for word variation.

Thus it becomes possible to cover completely all the types of word variation in Russian which is extremely important in view of the special position occupied by Russian in our system of automatic machine translation.

A simple mathematical calculation proves that on increasing the number of languages in a MT system, the practical necessity of using some languages as an intermediary language increases that is, an intermediate language into which translations are made in every case in order that further translation, if required, can be done from this language.

The most diverse opinions have been expressed as to which language should be used as an intermediary in automatic translation, in particular the idea that a special artificial language should be built up for this purpose, or several such languages.

The most rational alternative intermediary language seems to us to be a natural language, preferably that of the country in which the translation is made.

In the conditions in our country, translations of scientific-technical literature into and from Russian is most urgent. In all cases, even if the task were translation from English to German of a book not published in Russian, it would be very interesting and useful to translate such a book into Russian also. Consequently,

* See in greater detail paper by Lyu Yan Tsyuan' at conference on MT, Moscow, 1958.

† The content of these classes has been defined above.

every time a book was translated from one foreign language to another we could obtain a Russian version as a "by product" and this would increase further the advantages of machine translation over ordinary translation,

## *References*

[1] Machine translation of languages. N.Y., 1955.

[2] V. V. Vinogradov " Fundamental types of dictionary meaning of words". Voprozy Yazykoznaniya, 1954, 5.

[3] L. Wundheiler " Invariant prerequisites of all translations" (mimeograph).

[4] L. Bloomfield "Language ". N.Y., 1953.

[5] E. Reifler. Paper at conference on MT, 1952 (material of session).

[6] O. S. Kulagina and I. A. Mel'chuk. "Machine translation from French to Russian ". Voprozy Yazykoznaniya, 1956, 5.

[7] S. I. Ozhegov, "Three types of encyclopaedic dictionary of modern Russian". *Ibid.* 1952, No. 2.

[8] V. Oswald. Word-by-word translation (mimeograph).

[9] A. V. Fedorov. " Introduction to the theory of translation".

[10] M. M. Morozov. "Technique of translating ".

[11] S. S. Tolstoy. " Principles of translation from English into Russian".

[12] I. Bar-Hillel. " Can translation be mechanised?" Amer. Scientist, 1954, No. 42, pp. 248-260.

[13] A. I. Smirnitsky. " Theoretical course in modern English". Delivered to the Philology Faculty, MGU, 1949-50.