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

Session 3: CURRENT RESEARCH

MT RESEARCH AT THE UNIVERSITY OF CALIFORNIA, BERKELEY

Sydney M. Lamb
University of California, Berkeley

People sometimes ask me how we plan to instruct the machine to translate Russian sentences. I usually reply that questions of this kind remind me of the old Chinese recipe for tiger stew, which begins: "First you must catch the tiger".

No one can construct a program to translate Russian accurately without first obtaining the information that such a program must be based on; and the obtaining of that information requires a far more detailed analysis of Russian than has ever been made.

It is therefore a very curious thing that so much of the work in the field of Russian-to-English MT has been devoted to writing translation programs instead of investigating the structure of Russian. It is as if we had a staff of people trying to cook tiger stew without having caught the tiger, some of them using a pussy-cat instead as a result of mistaken identity, others in the belief that with a skillful enough job of cooking and with the proper exotic spices, the tiger meat will somehow magically get into the stew, thus making unnecessary the unpleasant task of actually going out to catch the beast.

It is surely not extreme to assert that about 90% of the job of developing a machine translation system is obtaining the necessary linguistic information, while the actual programming of the system, based on the knowledge obtained, amounts to 10% at most. It appears that in the past most of the research that has been done in Russian-to-English MT has been devoted to the 10%. The function which has been assumed by the group at the University of California, Berkeley, is to attack the other 90%.

The analogy to tiger stew, of course, is not perfect, since the idea that one can eventually get a suitable automatic translation system as a result of numerous refinements made upon an imperfect system is not as ridiculous as the notion that by making improvements on pussy-cat stew one will eventually end up with tiger stew. But the successive-approximation approach does have certain shortcomings. In the first place, when one has a trial-translation system, one never knows how much of the necessary linguistic information has already been obtained, and how much remains to be assembled. This is because it is in the nature of this approach to provide the trial-translation system with some means of dealing with everything that might come up, so that, in effect, whenever some situation occurs which cannot be handled with assurance of accuracy, a guess is made. In other words, it is not a part of the approach to try to make a distinction between what is already known and what remains to be discovered since each successive trial-translation system is set up in its entirety as an integrated hypothesis to be tested on a new text. This means that on that glorious future day when some text will have turned out to be correctly translated in its entirety, it will be presumed that all of the rules in the system are valid. But by the same token, since they also all have the same value at all times prior to that millennial occasion, the entire system (except to one who is intimately acquainted with its construction) is as weak as the weakest rule contained in it. And that is very weak indeed. This would not be the case if effort were made to distinguish those rules that are based on sound knowledge of the linguistic facts from those which are guesses formulated primarily on the basis of a few text examples. But in the outputs from trial-translation systems that I have seen, no such distinction is indicated.

Aside from the lack of precision inherent in such a system, it puts a needless burden on the investigators whose job it is to improve the trial translator by examining trial translations, because the weaknesses of such a system do not always reveal themselves in a trial translation. In fact, one might say that they <u>usually</u> fail to reveal themselves. There are two reasons for this circumstance. In the first place, downright errors can be made which are not identifiable as errors (except by careful comparison with the original text by one who knows the source language). But more important is the fact that even a bad rule makes a right guess a large percentage of the time. Indeed, a rule which makes a correct guess 95% of the time is still a bad rule. Thus, for each of those 95 out of 100 occurrences of the form affected by such a rule, its weakness will not make itself apparent in the output of the trial translator, simply because the correct guess was made.

In addition to the difficulty of locating the weaknesses, there is the drawback that, once located, the imperfect rules must be reworked and re-programmed. This means that every programming of a rule, except that glorious final correct one, will have been wasted effort.

For these reasons, among others, the Berkeley group, instead of spending its time programming a machine to make guesses, has embarked upon a direct and concerted effort to catch the tiger. Our project started officially in October, 1958, under the financial sponsorship of the National Science Foundation, but a certain amount of preparatory work had been done during the preceding months.

In setting out to catch the tiger, it is important to have a carefully formulated plan of attack. To have such a plan, one must know first of all just what it is that has to be obtained; in addition, the proper equipment with which to operate must be available; and finally, some procedure whereby the equipment is to be used to attain the goal must be worked out.

Our equipment is a set of concepts, principles, and techniques drawn largely from the general tradition of structural linguistics, but with some additions and modifications. Our goal is the capability to produce a particular type of translation, having certain clearly understood properties. Applying the conceptual equipment to the goal makes it possible to work out in broad outlines the design of a maximally efficient system for automatic translation, and from these broad outlines and the goal one may determine just what kind of knowledge about the structures of the languages concerned needs to be obtained. By conducting such planning in advance of large-scale accumulation and manipulation of data, it is possible for us to concentrate our efforts on the attainment of a single accurate and economical automatic translation system, dispensing with the construction of partially effective systems as intermediate steps.

To explain what we are doing, then, I must say a little bit on each of four areas: (1) The concepts for dealing with linguistic phenomena; (2) the kind of translation we want to have the system produce; (3) the general design of the projected automatic translator; (4) the kind of linguistic information needed and the means of obtaining it.

Session 3: CURRENT RESEARCH

Linguistic Concepts

There is time here only to mention briefly a few of the most basic linguistic concepts that are directly applicable to machine translation. In the first place, a distinction must be made between the *morphemic level* and the structural level of the *expression*, which in turn is to be distinguished from the non-structural or peripheral level of the expression. The expression has to do with the physical medium into which information is encoded for transmission from one user of the language to another. The two forms it may take in natural languages are speech and writing. In spoken language, the levels of the expression are the *phonemic* and the *phonetic*, of which the former is part of the linguistic structure, while the latter has to do with the actual speech sounds. The corresponding levels for a written language are the *graphemic* and the *graphetic.* Any text or portion of a text has simultaneous existence on all of the levels, but the various elements which are set up on each of the levels in the course of analysis are endowed with a separate existence by the grammarian. An element or combination of elements on a given level may be referred to as the *representation* on that level of the portion or portions of a text which it accounts for. We may also speak of it as representing an element or combination of elements of another level which accounts for the same portion or portions of a text. Thus the phonemic entity /t/ or the graphemic < ed> are representations of the past tense morpheme in such forms as "walked" (/wokt/).

The basic element of the graphemic level is the *grapheme*. It corresponds roughly to the letter, but there are some differences. For example, whereas among letters we have an entire set of upper case forms going with the lower case ones, after a graphemic analysis there will be a grapheme of capitalization and only one set of letters. On the graphetic level are the *graphs*, i. e. , the actual printed marks on a page, having many kinds of differences from one another of which only the distinctive ones are reflected on the graphemic level.

The fundamental unit on the morphemic level is the *morpheme*. The representation of a morpheme on the graphemic or phonemic level, as the case may be, is a *morph*, and if two or more morphs are alternate representations of the same morpheme, they may be referred to as its *allomorphs*. For example, in written English, the morpheme {child} has two allomorphs, namely < child> and <childr>, the latter occurring with the plural morpheme. That morpheme, in turn, has several allomorphs, including <s> as in "tigers", <es> as in "boxes", and <en> as in "oxen" and "children".

On any level there are units larger than the basic elements, these units being combinations of elements. Thus a morph is a combination of graphemes. A word is a sequence of graphemes which can occur between spaces. A morphemic word is the corresponding unit on the morphemic level. Combinations of elements and classes of elements always have their existence on the same level as those elements. The treatment of such combinations and classes is therefore concerned with the level to which those elements belong. This means that the notion of a syntactic level as being something different from the morphological or morphemic level is unacceptable in this system. Syntax is concerned with the morphemic level, since it has to do with material which is made up of morphemes. It is possible, however, to distinguish between morphology and syntax as two areas concerned with the morphemic level on the basis of different types of treatment being suited to inner-layer as opposed to outer-layer constructions. The traditional doctrine has been that combinations of morphemes to form words come under morphology, while syntax has to do with combinations of words. This division, however, has increasingly come to be recognized as an artificial one which gives entirely too much importance to the space (or, for the corresponding situation in spoken languages, to junctural phenomena), whose grammatical significance is actually only incidental. Some linguists have advocated doing away with the morphology-syntax distinction altogether. However, a strong case can be put forth for making the division between recurrent and non-recurrent constructions (with some qualifications). This would enlarge the area of syntax to include inflectional as well as productive derivational constructions. The items which serve as ultimate constituents for this kind of syntax may be called the lexemes. In a description of a language according to this system, the lexemes are the basic units of the lexicon as well as of the syntax. The reason for making the division in the manner indicated may be briefly explained in about two or three sentences. The use of the construction, as it is described below, to characterize combinations of linguistic units permits large bodies of data to be described in single small statements. These constructions should therefore be used throughout the entire area in which they are serviceable. The only part of the morphology-syntax area in which they do not contribute to efficiency of description is that which deals with the inner, non-productive layers of derivation, the description of which requires individual treatment of the various forms involved.

The representation of a lexeme on the graphemic (or phonemic) level may be called a lex. Thus a lex can consist of one morph or of more than one, and a word may consist of one or more lexes. Just as a morpheme can have allomorphs, so any lexeme containing such a morpheme has allolexes.

An element on a structural level has as its properties a definite relationship to items of adjoining levels, such as that between a morpheme and its allomorphs, and a distribution relative to other elements on its own level. On the basis of significant similarities in distribution, elements can be grouped into distribution classes.

The syntax of a language can be completely described by means of (1) a list of the distribution classes of lexemes, with the membership of each; and (2) a list of constructions. A construction is completely characterized by specification of (1) the distribution classes of the constituents which enter into it; (2) the order in which the constituents occur relative to each other; (3) the syntactic juncture (such as space or comma), if any, which occurs between the constituents; and (4) the distribution class of the constitutes. The only important difference between this view of the construction and most earlier ones is the fourth item named above, namely, specification of the distribution class of the constitutes. With this property included in the characterization of a construction, all distribution classes of units on the morphemic level which can occur as constituents are defined either by constructions or by the listing of lexeme classes.

According to the system under discussion, all of the semantic material is accounted for by the morphemes present, and is therefore predictable from them. In other words, none of the meaning is assigned to constructions, since to make such assignments is multiplying entities beyond necessity. Where differences in order of the same forms appear to be associated with differences in meaning, these meaning differences can be accounted for in terms of alternate meanings, or allosemes, of the morphemes involved. Such allosemes are in complementary distribution by virtue of the very fact that they are associated with differences in the order of the morphemes representing them.

Constructions generally cover numerous forms. Any specific occurrence of a form representing a particular construction may be called a constitute. Its immediate constituents are members of the classes indicated in the statement of the construction, and they may be called partners of each other.

Type of Translation Desired

The most important requirement for a translation of expository literature is accuracy. This means, above all, that extreme care must be taken to avoid allowing the machine to make guesses, even when they have a 95% chance of being correct. Wherever it is not possible to provide the machine with the means for furnishing correct target representations with assurance of accuracy, there should be a willingness to admit the fact, so that a choice can be offered between the alternatives which are possible.

The need for accuracy will also be served if the system produces translations which are as close as possible to the original text. Departure from the wording of the input text should be allowed only to the extent necessary to insure readability and intelligibility. It is therefore unnecessary, and maybe even undesirable, that the English translation conform in all respects to the rules of English style that would apply in the writing of an original text in English. To the extent that the flavor of the source language can be preserved in the translation without impairing readability and intelligibility, so much the better.

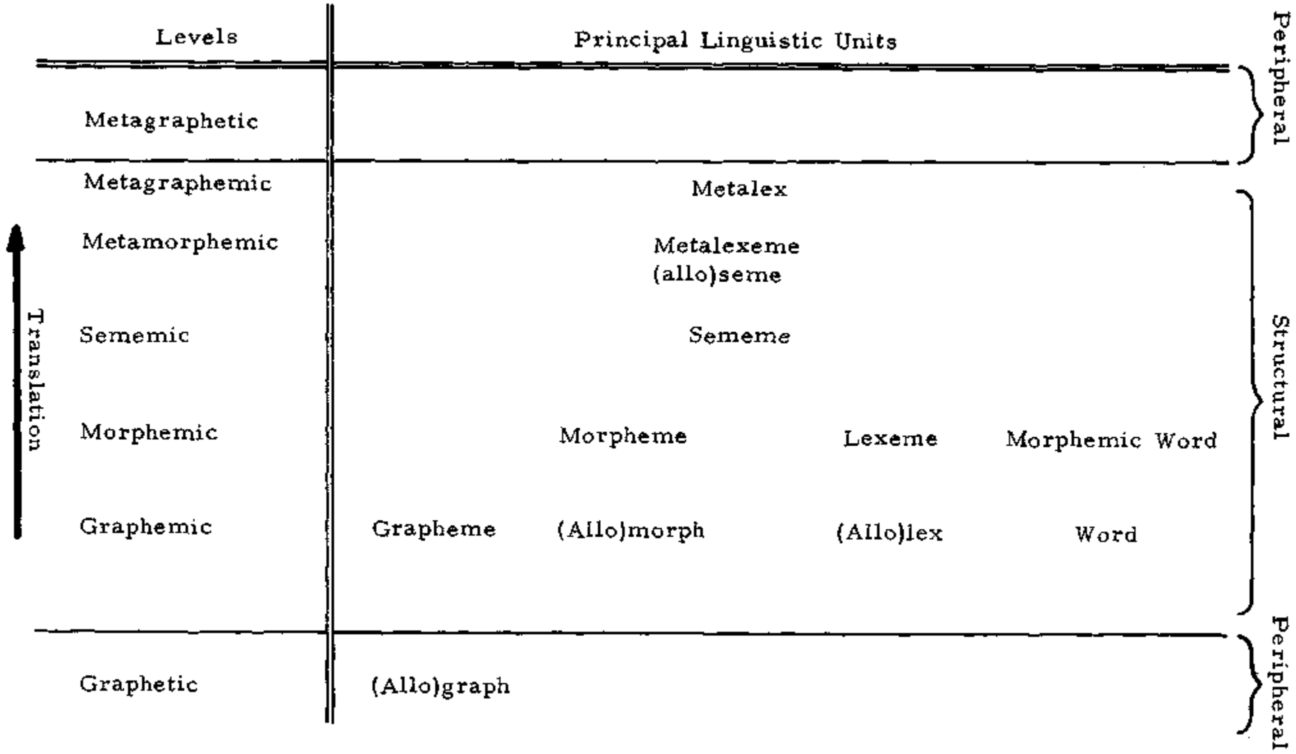
This principle provides a fairly precise means of defining the most suitable translation for any specific sentence. If we add to it the principle that a choice between alternate target representations is allowable wherever a single representation cannot be selected with assurance of accuracy on the basis of knowledge attainable within the foreseeable future, then we have a clear characterization of the goal which must be achieved. It is expected that the amount of knowledge attainable within the foreseeable future will be such that choices will have to be given on the average only about once or twice per sentence and that they will ordinarily be between only two alternatives.

Design of Translation System

Translation involves two languages, and linguistic structure can efficiently be treated in terms of levels. Therefore a translation system can be viewed as one integrated system of levels made up of the two systems back to back, as it were, and joined to each other at the middle by what we may call a sememic level (see Figure 1). In an integrated system designed for a single pair of languages, the nature of the sememic level will depend upon the semantic interrelationship of the two languages concerned. The levels corresponding to the morphemic and graphemic levels of the target language may be called the metamorphemic and metagraphemic levels, respectively, when incorporated into such an integrated system. (The Greek prefix meta- corresponds to the Latin trans- and is the prefix used in the Greek word for "translate", which is metapherein, of which the pher is the same root as the "fer" in "transfer".) In a translation system, then, the structural levels, in order, are graphemic, morphemic, sememic, metamorphemic, and metagraphemic; and the translation process can be viewed as a series of conversions of the linguistic material from one level to the next until the metagraphemic level is arrived at.

We must now consider what the basic units should be which are to be taken through this conversion process. People in the MT field have talked much of word-for-word translation and sentence-by-sentence translation. It is apparent that not all who have been concerned with this question have made a distinction between word-for-word translation and word-by-word translation, nor between sentence-for-sentence and sentence-by-sentence translation. It seems to have been widely assumed that a word-by-word translation is necessarily also a word-for-word translation, but this is by no means the case. The fact that translation is done on a word-by-word basis does not in any way preclude consideration of the environment, in all of its relevant aspects, during the process of translating a word; nor does it rule our arrangement of the target representations in an order different from that of the words they represent. In fact, translation

Figure 1 - LEVELS OF A TRANSLATION SYSTEM



Levels	Principal Linguistic Units					
Metagraphetic						Peripheral
Metagraphemic			Metalex			Structural
Metamorphemic			Metalexeme (allo)seme			
Sememic			Sememe			
Morphemic		Morpheme		Lexeme	Morphemic Word	
Graphemic	Grapheme	(Allo)morph		(Allo)lex	Word	
Graphetic	(Allo)graph					Peripheral

on the basis of units smaller than sentences, if done properly, takes into consideration more environment than sentence-for-sentence translation does, wherever necessary, and is therefore more effective in all situations in which consideration of material contained in a preceding sentence is necessary for a good translation. In other words, the only good translation is a text-for-text translation. But the sentence is not only too small a unit for x-for-x translation, it is also too large a unit for x-by-x translation. Efficiency demands that units much smaller than sentences, and, in fact, smaller in many cases than words, be used as the items to be converted from level to level. The most efficient items for this purpose are units which correspond closely, although not exactly, to the lexemes. Because of the general similarity, however, the term lexeme can be used for these units in discussing a translation system. The use of the lexeme (rather than the sentence) as the basic unit of translation makes it possible for the machine to avoid doing syntactic analysis of those features of sentences which can have no bearing on the translation, while allowing it to do all the analysis that is necessary for a good translation in every situation. To be both effective and efficient, then, translation should be neither sentence-for-sentence nor sentence-by-sentence, but lexeme-by-lexeme and text-for-text.

As in a description of a single language, the lexemes of a translation system are the basic units of the dictionary. Further details concerning these units are given in my paper "Segmentation", to be given in Session 7 of this symposium.

The representation of a lexeme on the sememic level may be called a sememe, and the representation of a sememe (or the lexeme it represents) on the metamorphemic level may be called a seme. If a sememe (or lexeme) can be represented by more than one seme under different circumstances, the alternate semes may be called its allosemes. A seme consists of zero or more metalexemes and specification of the order in which certain metalexemes are to be arranged.

Conversion from the graphetic to the graphemic level is done by the keypunch machine and operator (as long as character readers are not available), and the graphemic-to-morphemic conversion is accomplished primarily by the process of dictionary lookup, a process which must include segmentation of the graphemic material into lexes. But after the lookup the conversion to lexemes is still not complete for lexes which can represent more than one lexeme. For these, the determination of the correct lexeme requires consideration of the environment. Let us take, for example, the graphemic suffix -a. It can represent any of a number of entities on the morphemic level, such as nominative singular, genitive singular, nominative plural, or present gerund. Ordinarily, examination of the immediate environment, in this case the preceding stem, suffices to remove the ambiguity. (For some nouns, however, -a can represent either genitive singular or nominative plural, and a wider environment must be exmined.) Let us suppose that it is determined to be genitive singular; that is, it represents the two lexemes: genitive and singular. The genitive lexeme can be represented by any number of sememes, one of which is to be selected in each instance by examining the environment, according to the instructions which must be provided. Let us suppose that in our example the sememe chosen is "possessive". This sememe has two allosemes, [-'s] and [*of] where [*] indicates that the [of] is to be placed in front of the noun phrase on the metalexemic level. Another example of a sememe with allosemes is "comparative", which may be represented on the metalexemic level by either [more] in front of the adjective or the suffix [-er]. In general, however, since Russian and English, being related languages, have similar semological structures, the sememic level can be so set up that most sememes do not have allosemes. It is often convenient, therefore, to think in terms of converting directly from the morphemic to the metamorphemic level and to speak of the allosemes of a lexeme, without regard to the intervening sememes. Ordinarily the instructions for selecting the proper alloseme of a lexeme should be contained in the dictionary entry for that lexeme.

In working with relationships between the morphemic and metamorphemic levels, it is helpful to distinguish certain different types of possible situations. In the first place, allosemes may or may not contain specifications that certain metalexemes or groups of metalexemes are to be arranged in an order different from that of the lexemes they represent. For example, some allosemes of participial lexemes must contain a specification that the metalexemic representation of the entire participial phrase is to be placed after that of the noun modified. Wherever such specification is not given, the metalexemic material is to be arranged in the same order as the lexemes represented. Aside from the arrangement specifications, a lexeme may be represented by a single metalexeme, by zero (like most occurrences of case suffixes of adjectives), or by a combination of metalexemes. In addition, a combination of lexemes may be represented by a single metalexeme, e.g. , takim obrazom = "thus". Such a metalexeme may be called a portmanteau metalexeme. Or a combination of lexemes may be represented by a combination of metalexemes, as in nesmotrja na = "in spite of". Whether it contains one metalexeme or more than one, a seme which represents a combination of lexemes may be called a portmanteau seme.

While most of the operations which must be performed in translating a Russian text should be carried out according to compressed instructions contained in the dictionary entries for the lexemes concerned, there are certain types of information about syntactic structure which are needed so often in the course of selecting representations of lexemes that some syntactic analysis ought to be done for each portion of text as a whole, after conversion to the morphemic level, before the lexeme-by-lexeme selection of sememic representations begins. The most efficient method of obtaining this information is one which makes use of constructions as defined above and proceeds On a lexeme-by-lexeme basis, from left to right, making groupings of the lexemes into larger constituents. It makes use of a concept of relations between partners of a construction like that of the Copenhagen school, in which each partner either may or may not presuppose the existence of the other. From a description of the necessary syntactic information done in terms of constructions, a syntactic table can be made, access to which is to be had by direct addressing, using the distribution-class symbols of the lexemes as addresses. At the entry in the table for each lexeme the information is given as to what grouping that lexeme enters into, for any construction in which it presupposes the other member. In other words, the constructions are entered in the table under the peripheral constituent class (i. e. , the class which presupposes the other but is not presupposed by it). For purposes of this system, the peripheral constituent class of an exocentric construction is the one which occurs in the smallest number of constructions. For coordinate constructions, the first member may be selected as peripheral for these purposes, simply because the grouping process will proceed from left to right. The table entries for lexeme classes which are always nuclear (i.e., never peripheral) will instruct the machine to do nothing but move on to the next lexeme. For every construction in which the constitute class is the peripheral constituent class for a larger grouping, information pertaining to the larger grouping is to be provided in the same table entry.

The placement of grouping instructions under the peripheral (i.e., the least presupposed but most presupposing) constituent class guarantees maximum efficiency by keeping to an absolute minimum the searching for items in the environment which may not be there. Limitations of time preclude more detailed description of this method here.

Catching the Tiger

From an understanding of (1) the properties which a translation system should possess and (2) the characteristics of the ideal translation, as sketched above, it is possible to get a fairly clear picture of just what kind of knowledge needs to be obtained. It is obvious, first of all, that all of the available information which has resulted from previous research, and which is now recorded in numerous grammars and dictionaries and the like, must be called upon extensively, but critically. Valuable as it is, however, it is far from sufficient to provide the information needed for an automatic translator. In particular, the area between the morphemic and metamorphemic levels is for the most part a vast uncharted wilderness. A considerable amount of new knowledge, especially about the semological structure of Russian, must therefore be acquired. All the available means of obtaining the information must be exploited, especially analysis of texts. We hear a great deal about text analysis these days in the MT field, referring to types of operations which the professional linguist would hardly recognize as analysis in the familiar sense of the word. Let me make it clear that when I speak of analysis I am using the term in that old-fashioned, less-abused sense. I am not using it to refer to trial translation or postediting of trial translation or the like, but to a direct process of accumulating information.

Another way of stating the principle involved in lexeme-by-lexeme translation is to say that all features of a proper translation must be assignable to lexemes of the source language text. In conducting the analysis, therefore, it is necessary to find the proper place to which to assign each feature that is to be incorporated in translations. As has been stated above, the most faithful translation of expository literature is that which departs from the wording of the original only to the extent necessary to insure readability and intelligibility. Within this limitation, the translation should be as polished as possible. A translation having these characteristics may be called the preferred translation. With a few additional specifications it is possible to define the preferred translation for Russian scientific texts in such a way that, for most sentences, different properly trained analysts will arrive at either the same preferred translation or versions which differ only in features which are insignificant for purposes of the analysis system. One of the most important features of the analysis being conducted is the working out of preferred translations and the assignment of each feature thereof (including, of course, order changes) to individual lexemes of the Russian text. Further details of the system are given in a forthcoming paper.

Research Tools Available

In closing, I would like to mention some of the systems and materials which our group has been producing that will contribute to the future of MT. Each of the ten items listed below is believed to be either the best of its kind or the only one of its kind in the MT field, and any of them will be made available to qualified workers who make the proper arrangements with us.

1. A maximally effective segmentation system for Russian. Some of its characteristics are described in my paper to be given in Session 7 of this symposium.

2. Two grammar-coding systems based on this segmentation system. One of them is human-oriented, the other machine-oriented and more detailed. The latter is directly convertible into the former.

3. A dictionary, also based on our segmentation system, which has an estimated vocabulary coverage of 300, 000 graphemic words. Deficiencies of current knowledge are reflected in varying degrees of incompleteness of the information provided in the entries.

4. A dictionary system, for use on an IBM 704 or similar machine, which allows for a vocabulary coverage of up to 500, 000 graphemic words and accomplishes lookup and segmentation at a speed of over 7, 000 words per minute.

5. A system of graphemic coding of Russian scientific text for machines which accommodates the Cyrillic, Greek, and Latin alphabets, Arabic numerals, mathematical and chemical symbols, punctuation, italicization, capitalization, subscripts, and superscripts. The system is so designed that Russian text can be punched directly by a properly trained operator, without pre-editing (other than the assignment of location reference numbers).

6. A system for analyzing Russian scientific texts to obtain information needed for an automatic translator. This system is described in a forthcoming report.

7. An analysis, according to this system, of about 30, 000 words of Russian texts in the field of biochemistry. An additional 30, 000 words of text have been put on punched cards for analysis by our automatic text analyzer, which is still under construction.

8. A program for obtaining various kinds of information from the analyzed text by the use of an IBM 704 computer.

9. A catalog of types of situations in which metalexemic material must be arranged in an order different from that of the lexemes represented, for the sake of readability or intelligibility. The tabulation is now believed to be nearly complete, and the situations have been classified and assigned to associated lexemes. The associated lexemes for an order-change situation are those which have been selected as constituting a small unified class whose presence (1) is a necessary condition for the occurrence of the situation, and (2) is not presupposed by any other class of lexemes whose presence is also a necessary condition for the occurrence of the situation.

10. A method (described above) for automatic syntactic analysis of Russian text, and information on Russian syntax with which the method is to be implemented. The necessary syntactic information is still being compiled, but the material now available is perhaps sufficient to be of some interest to other workers.