[*Mechanical Translation*, vol.4, no.3, December 1957; pp. 70-75]

Contextual Analysis

Kenneth E. Harper, University of California, Los Angeles, California

Ambiguity, both syntactic and semantic, a problem that arises in the translation of Russian to English because of polysemantic forms in Russian, can be resolved by an analysis of the context in which the polysemantic form occurs. This requires a systematic study of context so that word classes which determine the value of ambiguous forms can be established.

IN THE VARIOUS PROPOSALS for word-for-word machine translation of Russian scientific literature into English, each word in the sentence is considered as a separate entity. If a word has more than one English equivalent, or more than one possible syntactic value, the alternatives must be listed. The chief difficulty with the resulting translation is its prolixity: the reader finds himself confronted with numerous alternatives, both syntactic and semantic, in every sentence. The extent of the problem of ambiguity is suggested by the following figures: from a sample Russian scientific text, 43% of the running words were found to be polysemantic (this in addition to syntactic ambiguities which the reader must solve on the basis of numerous alternatives given him in every sentence).

Context

The difficulty with word-for-word translation, then, is that it is really "words-for-word translation". [1] The solution to the problem lies in the reduction of the number of choices confronting the reader by the mechanical selection of the proper (or actual) syntactical and semantic equivalent from the various potential equivalents. Obviously, the solution can be attempted along lines as infinitely complex as those involved in "human translation", in which judgments are based on "context", experience and even upon "taste". Of these the element of "context" is, to some degree, determinable by mechanical means. In its general sense, context signifies environment, i.e., surrounding words in a sentence, surrounding sentences and paragraphs, extending to the broad category of subject areas. The question arises: Is some more limited use of context analysis possible in MT, and how effective is such analysis in the removal of ambiguity?

In an attempt to answer this question, the potentialities of a "contextual analysis" of each ambiguous word (syntactically or semantically ambiguous) have been studied, such analysis to be limited to immediately contiguous words. Thus, for a given ambiguous word (x), reference may be made to the preceding word (x-1) or to the following word (x+1). (In specified instances, reference may be made to words which are separated by neutral words from (x) word.)

The value of this limited contextual analysis was suggested by the inflectional nature of the Russian language. For example, the English preposition, 'of, indicating possession, does not have a "word equivalent" in Russian; the 'of' is generated by the genitive case of the noun or pronoun (<u>добавление смеси</u> = 'the addition <u>of</u> the mixture'). Two difficulties arise

---

1. The problem of word order is not critical in MT, particularly for technical material. Even in the general literary language, the word order, subject-verb-direct object, is preserved in 85 - 90% of all sentences (according to a study of 5000 pages of Russian prose text, cited in <u>Voprosy grammaticheskogo stroya,</u> Izdatel'stvo Akademii Nauk SSSR, Moscow, 1955, p. 471).

in straight word-for-word MT: 1) difficulty of identifying the genitive ending for most nouns, so that the above Russian words may theoretically mean 'the addition to the mixture', 'the addition the mixture', or 'the addition the mixtures', as well as the translation given above; 2) the 'of generated by the genitive case is often disregarded, under the condition, for example, that the word is preceded by a preposition which governs the genitive case. The task of deciding whether or not to retain the 'of' falls upon the reader. The problem results, of course, from the syntactical compactness of inflected languages. Since syntactical information in Russian is contained not in discrete items (individual inflected words), but in the relationship between words, a comparison process is imperative.

A second reason for believing in the potential of contextual analysis is the effect that consideration of immediately contiguous words has upon the removal of semantic ambiguity of a given word. Professor Kaplan's study on this problem suggests that a marked reduction of ambiguity is the result of considering one or two words preceding and following the ambiguous English word.[2] This is a completely-virgin field of investigation, but preliminary studies indicate that within a closed area of discourse, such as Russian technical literature, the problem of multiple meaning can be satisfactorily handled through the analysis of contiguous words.

In the two following sections studies on the effect of syntactic and semantic clarification by this method are summarized.

Clarification of Syntax

It is essential in this system that any given word in a Russian sentence be subject to retention and further inspection; in other words, location of the item in the memory is only (or nay be only) half the job. Even after its grammatical features have been determined, whether in a paradigm or stem-affix machine dictionary, :he word is not to be printed by the output den.ce until a "go ahead" signal is given. In theory, every word in a sentence is potentially useful to a contiguous word; every word is a potential determiner, and, if it is in any way ambiguous, a potential determinee. Our problem is to discover the manner in which this relationship is expressed, and to represent it in codable form. In certain instances, as in the relationship between adjective and noun, for example, the mutual influence is recognizable in terms of conventional grammar; more frequently, the relationship is unpredictable and must be discovered by observation of behavior in a large number of situations. In any event, the ability to make reference to words in immediate contiguity is inherent to this system.

For purposes of syntactical clarification, conventional grammatical concepts are quite useful. It is helpful, for instance, to have available, in coded form, the following information for words in a Russian sentence: part of speech of all words; case, number, and gender of nouns; the infinitive form and tense of verbs; case and number of certain adjectives, etc. Reference to this information may be helpful in contextual analysis. It should be stressed that reference is made to these coded features, rather than to "the word" itself. In the latter process, we become involved in the identification of idioms, i.e., in the problem of lexical relationship; our present interest is in the structural relationship and its effect upon clarification of syntax.

The processing of syntactically ambiguous words may be summarized in the following descriptive terms:

1) Nouns

a) Genitive Case

For masculine nouns, this case is identifiable by ending (disregarding, in technical Russian, the almost non-existent animate noun). For all neuter and feminine nouns, this case is ambiguous by ending in the singular. For all unmodified nouns which are definitely or potentially genitive case, by ending, the English preposition 'of' is generated only under the condition that the preceding word is a noun. The 'of' is to precede the noun identified as genitive; if adjectives precede the noun in question, the 'of' is to precede all such modifiers. In referring to the part of speech of the preceding word, modifiers of the word in question are ignored.

добавление смеси
= 'addition (of) the mixture'

добавление этой смеси
= 'addition (of) this mixture'

2. Kaplan, Abraham, "An Experimental Study of Ambiguity and Context", Mechanical Translation, vol. 2, no. 2, pp. 39-46, November 1956.

The result of the above restriction (that the preceding word must be a noun) automatically eliminates the generation of the 'of' in the frequent instances where the genitive case is required by Russian grammatical rules, but where its identification only serves to hinder the translation, — for example, when the preceding word is: a preposition, a cardinal number, a comparative adjective, a negative (нет), a verb which governs the genitive case, words of quantity (много, сколько), negated verb, etc.

This rule, formulated purely on the basis of observed behavior, very accurately approximates the control over "context" unconsciously enjoyed by the human reader of Russian.

b) Instrumental Case

This case is not ambiguous by ending. Nouns in this case (and any preceding modifiers) are to be preceded by the English word 'by' ('with' in certain specified cases), except when the preceding word is a preposition, or a verb governing the instrumental case (which may also follow the noun).

c) Dative Case

This case may be ignored, since the generation of the English 'to' can be most economically handled in the dictionary listing of the manageable number of words which precede nouns in this case.

d) Nominative, Accusative, and Prepositional Cases

These may be ignored because of the factor of word order.

e) Number in Nouns

The plural number of all nouns is unambiguous, with the exception of neuter and feminine nouns in the nominative and accusative plural (where they are identical with the genitive singular). If these ambiguous forms have been identified as genitive (under la above), they may be automatically identified as singular also. In all other instances, the number of such forms can be satisfactorily determined by reference to the preceding word. The adjective and (in almost all instances) the preposition are absolute determiners of number; other forms which require the noun in the genitive case may also be utilized to determine the singular number of the ambiguous form (in instances where the English 'of' is not generated); the absence of these conditions, or the presence of a period or a comma in the preceding position, may be taken as an indication that the form is plural in number.

2) Adjectives

Often adjectives are useful in determining the case and number of nouns; otherwise, they may be ignored as to agreement with noun.

a) Short adjectives, singular, (in -zero, -а, -о) are to be preceded by the word "(is)" in translation; short adjectives, plural, (in -ы or-и) are to be preceded by "(are)". These English words are, further, to precede an adverb which may precede the short adjective.

Если температура очень высока
If the temperature (is) very high

b) Comparative adjectives: the word 'than' will be inserted in the translation if the following word is a noun.

3) Adverbs

The distinction between a short neuter adjective and an adverb is apparently impossible to make, since the forms are identical. Preliminary investigation shows that a high degree of accuracy can be attained by reference to context: if the following word is a modifier or a verb in the indicative, the word in question is an adverb; if the following word is an infinitive, the word in question is a short adjective. The accuracy of prediction can be increased by further extension of the comparison process. It is, however, doubtful that such refinement is necessary.

4) Participles

A participle may serve in a sentence as an "adjective", as a true participle or (rarely) as a noun. The decision as to its function in a given sentence cannot be made on the basis of form. Observation of its behavior, however, leads to the following formulation:

a) An active participle can be adequately translated as '-ing' Определяющий = 'determining'; a passive participle can be translated as '-ed' (определенный = 'determined').

b) If the participle agrees in case and number with the following word (a noun, or adjective + noun), it is treated as an adjective (i.e., as a modifier), число заряженных частиц = 'the number of charged particles' (rather than 'the number charged of particles').

c) If the participle does not agree with the following word, it is a true participle, число, определенное этим методом = 'the number, determined by this method.'

Again, although this formulation is completely arbitrary, no exceptions to its correct-

ness have been observed in a study of 132 occurrences. (Slightly less accurate results can be obtained merely by reference to punctuation: a preceding comma makes the word in question a true participle.)

The above represents the classes of syntactical problems which are encountered most frequently in Russian text. By application of well-defined rules involving reference to pre- or post-words, clarification can be attained to a very high degree of accuracy. A few minor problems remain, caused chiefly by "awkward" word order, inverted clauses, etc.

Conclusion: Syntactical ambiguity can be removed to a highly satisfactory degree by the comparison of ambiguous words with words in immediate contiguity.

Clarification of Semantic Ambiguity

It is obvious that problems of syntax and semantics are closely related. For purposes of discussion the two have been separated, and the latter has been arbitrarily divided into two categories: "structural" and "non-structural" clarification.

1. The most common instance of structural clarification is the determination of English equivalents by means of the grammatical case of contiguous words. Thus, the Russian preposition £ is translated as 'with' when the following noun is in the instrumental case, and as 'from' when the noun is in the genitive case. The English equivalent of other prepositions also varies with the grammatical case of the object, as set forth in dictionaries and grammars. These relationships are predictable and easily recognizable.

Behavioral analysis brings to light a great number of unsuspected semantic relationships between words of multiple meaning. These relationships have been only partially uncovered, but the semantic clarification so provided holds great promise in MT. An example is found in the Russian conjunction, и_, which is listed in dictionaries as: 'and', 'but', 'even', and 'also'. A test case was made of this frequent and annoying conjunction, on the assumption that perhaps its meaning could be determined by immediately contiguous words. On the basis of 200 occurrences in scientific text, it was found to be equated with the English 'and' whenever the preceding word was a noun (which situation prevailed in 70% of the total occurrences). By a slight extension of this comparison to other parts of speech and to punctuation, we can predict the correct equivalent of и in 90% of its occurrences.

Other examples of structural clarification of this kind include:

a) The word их. which serves in Russian both as a pronoun and pronoun-adjective ('them' and 'their' in English). It has been found that this word can be equated with the proper English word according to the nature of the following word (noun or non-noun).

b) Words which serve both as an adjective and as a noun, and whose English equivalent varies accordingly. Thus, данные is equated with ' given" when it is singular in number or when it agrees as a modifier with the following noun; in all other instances it is translated as 'data'.

2. "Non-Structural Clarification". Words of multiple meaning for which clarification by structural means is impossible constitute approximately one-third of the running words in a text. (This figure is in addition to idioms, which are a special problem.) In pursuit of the ideal — to select, within practical limits, a single correct equivalent for these words — we must look for some kind of contextual aid other than that supplied by grammatical features of surrounding words.

In the first place, it is clear that new techniques of lexicography for MT need to be developed. Reliance upon dictionary equivalents must be replaced by observation of the behavior of ambiguous words in given fields of technical writing. For example, if observation shows that the Russian изменение may be always equated with the English 'change', in texts on physics or mathematics, the nine equally possible dictionary variants ('alteration', 'fluctuation', 'variation', etc.) may be disregarded. Limited observation indicates that 'property' may be taken as the correct equivalent of свойство in the same field (as opposed to 12 dictionary listings); 'study' for исследование (7 listings); 'substance' for вещество (7 listings); 'body' for тело (8 listings); 'magnitude' for величина (15 listings), etc. In addition, superior techniques must be perfected for choosing the best "cover-word" from among a group of relatively synonymous equivalents. Existing "technical" dictionaries are in no sense idioglossaries, since they list a great variety of potential equivalents for most

words. A true idioglossary must be based upon the observed values of multiple-meaning words, with the emphasis placed upon singularity, rather than upon plurality, of meanings.

Regardless of the size of the context-sample, we must be able to observe ambiguous words in action: the kinds of nouns which follow certain prepositions, the kinds of adjectives which impart specific values to certain nouns, etc. An empirical study of this scope, practicable only with the aid of modern machine techniques, will go far towards unveiling the mysteries of "context". We have long since passed the stage in MT research when we should be bound by speculation of what "might be"; we need to take a bold step forward to find what actually exists.

The application of contextual analysis offers great potentialities for semantic clarification. In this instance, comparison of ambiguous words is effected with contiguous word classes. Word classes are simply groups of words (usually of like parts of speech) which have the common property of causing other words to behave in a predictable manner. For example, the Russian preposition по has ten potential equivalents when followed by a noun in the dative case; by reference to pre-determined noun classes we can reduce the number of choices to one, in most instances. (If the noun-object is an animate noun, по acquires the meaning, 'according to'; if the object is a verbally derived noun, the meaning is 'in'; if the object implies a path or a surface, the meaning is 'along'.) An extended survey of physics texts indicates that the vast majority of noun-objects after this preposition fall in one of these three classes. The word classes are formed purely on the basis of observed behavior; with further refinement and extension of research, it appears feasible that pinpointing of meaning will be possible for most occurrences of this most difficult preposition. Like procedures can be instituted for a great variety of ambiguous words.

The great advantage of using word classes is that the necessity of treating each new combination as an "idiom" is eliminated. It is apparently in some such fashion that the human translator chooses a particular equivalent for a given ambiguous word when he encounters the word in a novel or unremembered combination. In idioms, of course, the factor of memory proceeding from previous acquaintance with the combination, is essential. But when the human encounters the combination по оси for the first time, on what basis does he equate по with 'along' (the axis), rather than with 'in', 'according to', etc. ? It is possible that in some instances the human engages in a process of elimination, discarding from consideration certain inappropriate equivalents; it is also possible that the choice is often made purely on the basis of the "class" of noun-object (i.e., "axis" is associated with a class of words, including "line", "radius", etc., which is known, on the basis of previous experience, to impart the meaning 'along' to the preceding preposition). Just how decisive this type of word class association may be in the determination of meaning, and the extent to which the crudely formed classes described in the foregoing paragraph will answer the purpose, remains to be proved. It can safely be predicted that this kind of "contextual analysis" will be quite effective, particularly within specified areas of discourse.

Another type of ambiguity is posed by words which bear multiple meanings even within a specific area of discourse. The Russian noun напряжение, e.g., may be translated as 'tension', 'stress', or 'voltage'; it is obvious that any of these meanings may be applicable in a text on physics. A partial solution to the problem of choosing the correct equivalent may be sought in further refinement of the idioglossary: thus, in texts concerning electricity, 'voltage' may be predicted. The human translator often chooses 'voltage' because of the contextual aid provided by the subject area: specifically, he identifies the subject area by the title or beginning sentences of the text. Two mechanical methods may be adapted for determining the appropriate equivalents. One involves the employment of sub-idioglossaries (e.g., for the field of acoustics), — which may necessitate pre-editing, in texts which are not clearly or mechanically identifiable by subject area. Another possibility is the reference of multiple-valued words to certain key-words in the title or first sentences of the text. Preliminary study indicates that this approach may lead to unexpectedly positive results. To take an extreme example, it may turn out that the very presence of the word "polymorphic" in a title will fix the specific equivalent of the following polysemantic words in the succeeding text:

чистый	'pure', rather than 'clean', 'clear', 'net', 'smooth', 'absolute', etc.
твердый	'solid', rather than 'hard', 'tough', 'durable', 'stable', etc.
вещество	'substance', rather than 'matter', 'material', 'agent', 'composition', etc.
соединение	'compound', rather than 'fusion', 'connection', 'union', 'contact', etc.

(It should be noted that the fact that these words appear in an article on chemistry does not guarantee the same selection.) There may be no apparent reason that this selection of equivalents should be valid, and it is certainly possible to invent contexts within chemical literature where they would not be so. But, if on the basis of observation these equivalents are found to be adequate, there is a strong argument that the empirical evidence should be accepted and utilized.

There are, of course, words for which semantic clarification cannot be obtained by use of an idioglossary; the referent is not the subject area, but perhaps a contiguous word — an adjective for a noun, or a noun object for a verb. It remains to be seen whether or not the contextual aid provided by such contiguous words can be programmed in a non-idiomatic fashion, — i.e., not on a one-to-one basis. The goal should be the establishment of word classes of the "determining" words which will enable us to fix the semantic values of the "determinees".

The result of the aggregate of structural comparisons of this kind, and of the kind described in the preceding section, is, in effect, a new grammar — a structural, or analytic, grammar designed for the specific purposes of MT. There is no question that this approach, based on an analysis of ambiguous words in terms of coded features of contiguous words, is adequate for MT and is superior to the approach of conventional grammatical analysis.

From the point of view of methodology it is notable that a completely unexpected relation is found to exist between structural context and meaning. It should be stressed that the existence of this particular relationship has never been even remotely considered by Russian philologists. The connection is, of course, not absolute; it is merely one of the phenomena of language which can be discovered by observation, and which is sufficiently reliable to be of use in MT.

Conclusion: The value of contextual analysis for purposes of syntactic and semantic clarification should be evident. The plain fact, however, is that no systematic and thorough study of context has ever been attempted for any language. There is an overwhelming and immediate need for such a study, conducted over the range of a million or more running words in the scientific literature of a given language, with the help of machine techniques. The information and experience gained in such a study will be of great value for similar studies in other languages. Since our primary concern here is the behavior of words in context, the machine run should be constructed so as to give the researcher rapid access to numerous occurrences of ambiguous words in "real-life" situations. In line with Kaplan's suggestion, it may prove that five-word blocks (with the ambiguous word in the middle position) will be sufficiently large to establish semantic clarity and an adequate judgment of the effect of contiguous words.