Mechanical Translation and the Problem of Multiple Meaning †

A. Koutsoudas and R. Korfhage, Willow Run Laboratories, University of Michigan

THE UNIVERSITY OF MICHIGAN undertook research, late in 1955, in the analysis of language structure for mechanical translation. Emphasis was placed on the use of the contextual structure of the sentence as a means of reducing ambiguity and on the formulation of a set of operative rules which an electronic computer could use for automatically translating Russian texts into English. This is a preliminary report on the latter phase of the problem, stating the results and suggesting a practical method for handling idioms and the problem of multiple meanings.

It was decided that the first work would be done on Russian texts in physics, both because of the interest in this field and because of the general availability of texts. Some work has already been done in this field.[1] If this work proves successful, it will form a basis for work in other scientific, technical, and military fields.

A text was selected from a Russian journal on experimental and theoretical physics.[2] It was chosen to present most of the expected difficulties; i.e., stylistic, orthographical, grammatical, etc. On the basis of this text a vocabulary was set up and fifteen rules were established. (Subsequent work has altered the rules slightly to remove such obvious faults as the occurrence of "the" before proper names.) It should be realized, of course, that neither the vocabulary nor the rules were in generally applicable form. The vocabulary was simplified by applying a "one form, one meaning" rule whenever possible. Thus, inflectional endings were stripped from most word stems although in some cases a word was listed with two or three specific endings. Most words were given their scientific meaning only. Some words, however, occurred in more than one sense, or

were combined with others to form idioms; in which case more than one meaning had to be listed. Finally, the words were listed in conventional grammatical categories; i.e., verb, noun, adjective, etc.

In the long run, we expect that the concept of conventional categories will be completely abandoned. What we hope to have, instead, are word groups the interaction of which will provide the grammatical and syntactical information needed. The need for such grouping has been made apparent.[3]

The rules were developed empirically by analysis of the essential processes undertaken by a human mind in translating a foreign text. It was found that most of the rules involved either word order or the grammatical functions which in Russian are indicated only by case endings and which in English might be classified by inserting a preposition. In most cases the rules concerning word order were sufficient to eliminate the necessity of referring to endings. To test the adequacy of the rules, several volunteers who had no knowledge of Russian were asked to translate the original text, using only our rules and vocabulary.* Except for random, minor stylistic faults, it turned out that the resulting translations were clear and accurate. Being convinced that the rules are as complete as is practicable for the text, we are currently enlarging the vocabulary in preparation for future tests on different texts.

Perhaps the most significant result thus far is the success in handling multiple meanings,

† The work upon which this paper is based was performed under the Department of the Army, contract No. DA-36-039-sc-52654.

3. See V.H. Yngve, "Sentence for Sentence Translation", MT. Vol.2, No.2, Nov., 1955.

* The Russian text with the vocabulary and rules based on this text will be found on pp.48 to 49. A standard translation and a translation made with the help of the rules by a volunteer who had no knowledge of Russian are on pp.50 to 51.

1. See K.E. Harper, "A Preliminary Study of Russian". Machine Translation of Languages, The Technology Press of the Mass. Institute of Technology and John Wiley & Sons, Inc., New York, 1955.

2. Zhurnal Eksperimental'noi I Teoretichesk'oi Fiziki. Vol.26, No.2, pp. 189-207, Feb., 1955.

which has given us an insight into the problem of idioms. Although the problem of ambiguity as exemplified by this situation was greatly reduced by the use of a highly specialized vocabulary, the situation still occurred and a means for solving it had to be found. Published results on this problem have, generally, involved either a post-editor or a separate idiom dictionary.[4] These methods seem undesirable particularly in view of the additional computer time required for translation. Consequently, a method was developed which, it is felt, is widely applicable. The assumption was made that the specific meaning of a word could be determined from its context. It developed that not only is this assumption valid, but in fact we need not consider sequences of more than four words. The method used is the following:

All possible meanings of a word are listed, consecutively, in the order (1), (2),...........(n). In general, in order to have corresponding meanings mesh, it will be necessary to list some meanings for each word more than once, and to include some blank translations. When a word with multiple meanings is encountered, the number (n) of meanings is noted and translation is postponed. Subsequent words are examined for the number of possible meanings of each, until a word (X) with a single meaning is encountered. If there is only one word in the sequence preceding X, then the first listed meaning is assigned to this word. If there is more than one word in the sequence preceding X, we determine (M), the minimum of all (n) noted in the sequence. Let us denote by (i) [A] the i-th meaning of a word A, and by $\underline{0}$ a blank (null) translation.

Given a two-word sequence, A B, we consider (M) [A] and (M) [B]. If neither of these are blank, we translate, assigning meaning (M) to each word. If either of these is blank, we consider (M-l) [A] and (M-l) [B] and apply the same test to these. In this way, we find the highest numbered meaning which is not blank for either A or B and assign this meaning to each.

Given a three-word sequence, ABC, we consider (M) [B]. If (M) [B] is $\underline{0}$, we consider successively meanings M-l, M-2,....., as above, and assign finally to all three words the highest numbered meaning which is non-blank for all. If (M) [B] is not $\underline{0}$, then if (M) [A] and (M) [C]

are both $\underline{0}$, we assign meaning (M) to the three words; otherwise we search meanings M-l, M-2,......of all three words, applying the above rule.

In a four-word sequence, ABCD, (M) [B] is again considered. The procedure followed is that used for a three-word sequence, except that (M) [D] must be considered along with (M) [A] and (M) [C] .

In all cases, if no translation is found by the above procedure, we assign to each word meaning (1).

By properly ordering the meanings for each word (listing some meanings several times if necessary), it has been found possible to obtain valid translations for over 96% of the two-word sequences [The two exceptions which occurred, по делу and цель в, were easily handled by separately listing дел in the form делу , and цел in the form цель .] and for over 90% of the three-word sequences which might occur. These figures are based on the possible sequences without reference to their relative frequency of occurrence in actual use. It is not known how the difficulties in "properly" ordering the meanings will multiply as the vocabulary is increased. With each new word (or meaning) added, the order of the meanings previously listed may have to be changed *so* as to maintain consistency as much as possible.

In this system an idiom is handled as merely an additional meaning which is possible. A study of the structure of three-word idioms showed that generally the second word had the least number of meanings. On this basis it was decided to assign to the second word the entire idiomatic meaning, and to supply corresponding $\underline{0}$ translations for the other two words. Thus, for example, the Russian idiom по сути дела ("actually") would appear as по = $\underline{0}$, сут = actually, дел = $\underline{0}$. (Note the dropped inflectional endings.)

To illustrate this method, let us consider the eight Russian words том, дел, сут, цел, по, в, о, and теори. From these eight words it is possible to form 56 two-word sequences and 336 three-word sequences. However, of these only 29 two-word and 106 three-word sequences are linguistically possible. It is assumed, of course, that the appropriate inflectional endings are supplied in each case. (The list of sequences, with translations, is available on request.) By working with these 135 sequences it was found that the arrangement of meanings given in Table I is the best possible. There seem to be no algorithms for ordering the meanings, other than that the idiomatic meaning, if any, be

4. See, for example: "The Treatment of Idioms" by Y. Bar-Hillel, typewritten, 8 pages; "A Study for the Design of an Automatic Dictionary" by A.G. Oettinger, doctoral thesis, Harvard University, 1954.

the last <u>meaning</u> listed for at least one of the
words.

TABLE I

том	сут	теори	цел	в	по	о	дел
1. that 2. 0	1. essence 2. actually	1. theory 2. 0 3. theory	1. purpose 2. 0 3. 0 4. order to 5. target	1. in 2. 0 3. in 4. in 5. 0	1. by 2. 0 3. accord- ing to 4. 0 5. at	1. about 2. 0	1. fact 2. 0

It may be noted that on the basis of only the
three words по, сут, and дел the shorter ar-
rangement of meanings given in Table II suffices,

Table II

по	сут	дел
(1) by (2) 0	(1) essence (2) actually	(1) fact (2) 0

It will be observed that there is a certain

amount of redundancy inherent in this system.
However, it is felt that this is a minor fault;
first, because the percentage of redundant mean-
ings in the entire vocabulary appears to be
small (around five per cent) and second, because
this plan does not require a separate idiom dic-
tionary or other special devices which tend to
increase computer translation time. Although
further research is necessary for the complete
development of this method, we believe that the
theory used is valid and that it eventually will
lead us to the solution of most multiple-meaning
problems.

VOCABULARY AND RULES

NOUNS

```
Буссин -  Boussinet
врем - time
времен -(1) time (2) the period
вычитани - subtraction
движени - movement
действительност - reality
дело - (1) fact, (2) 0
значени - value
значениями - values
интервал - interval
корреляци - correlation
Крутков - Kroutkov
малост - shortness
момент - instant
некоррелированност - uncorrelativity
обобщени - generalization
Орнштейн - Ornshtein
основани - reason
Планк - Plank
последействи - after-effect
```

```
предполозкени  - assumption
промехутк      - interval
приращени      - increment
приращений     - Increments
процесс        - process
работ          - work
рассмотрени    - examination
результат      - result
результатам    - results
релаксаци      - relaxation
сил            - force
скорост        - velocity
создали        - (1) formulation
                 (2) formulate
сравнени       - (1) comparison
                 (2) as compared
Стокс          - Stokes
сут            - (1) essence
                 (2) actually
теори          - (1) theory
```

                                                        случайн -   random
                                                        справедлив -   correct
                                                        сравним -   comparable
                        (2) on the theory   том -   (1) that (2) 0
                        (3) in the theory   указанн -   indicated
течени   - (1) course   упорядоченн -   correlated
               (2) during the   физическ -   physical
удар   - collision
уравнена   - equation
ускорени   - acceleration   ADVERBS
Фоккер   - Fokker
формул   - formula   более -   a more
формулой   - by the formula   больше —   more
функци   - function   всё-таки - nevertheless
цел   - (1) purpose   достаточно - sufficiently
               (2) in order to   правильно -   correctly
частиц   - particle   после -   after
частот   - frequency   поэтому -   therefore
частност   - (1) particularity   соотвественно - accordingly
               (2) in particular   статистически - statistically
Эйнштейн   - Einstein   также -   also
                                                        точнее -   more precisely
                                                        учитывая -   by taking into
              VERBS   account

был — а — was
был — и — were   MINOR PARTS OF SPEECH
выражать - to express
оказыва - ется - proves to be   а - and
описыва - ет - describes   в - (1) in, (2) 0, *(3)* 0
отсутству - ет - is absent   даже - even
предполага - лась - was assumed to be   для -   for
предполага - лись -were assumed to be   если - if
привед - ет - will lead   и -   and
создать - to formulate   к -   to
явля - ется - is   когда - when
                                                        лишь - only
                                                        между - between
           ADJECTIVES   не -   not
                                                        но -   but
больш -   large   о -   (1) about, (2) 0
броуновск - Brownian   однако - however
выражающ - expressed   по -   (1) by, (2) 0
гидродинамическ - hydromatic   порядка -   within
законн - legitimate   при -   at
корреляционн - correlated   с(о) -   with
мал -   small   также -   also
марковск -   Markov's   то -   then
меньш -   smaller   что -   (1) that, .(2) that
небольш - small   этому -   <u>0</u>
независим - independent
некоррелированн - uncorrelated
несправедлив - incorrect   ABBREVIATIONS
неупорядоченн - random
остающ -   remaining   ДР — others
перв -   first   см — see
подобн -   such   т.е. - i.e.
полн -   complete
пригодн -   applicable
применим -   applicable   PRONOUNS
протекакщ - taking place
различн -   various   её - its
рассматриваемым - observed   она - it
сделан -   made

RUSSIAN TEXT

В первых работах по теории броунов-
ского движения /[1]/ (см. также /[22]/)
значения скорости частицы в различные
моменты времени предполагались по сути
дела статистически независимыми.
Соответственно этому была применима
формула Эйнштейна

$$M (x - x_0)^2 = 2 \qquad (1)$$

а также уравнение Эйнштейна-Фоккера-
-Планка, справедливое для марковских
процессов. В действительности, однако,
корреляция между значениями скорости
отсутствует лишь при достаточно боль-
ших интервалах времени между рассматри-
ваемыми моментами. Поэтому формула
(1) оказывается несправедливой для ма-
лых интервалов времени (порядка времени
корреляции для скорости).
В целях создания более полной теории,
пригодной для меньших интервалов вре-
мени, были сделаны предположения
(Орнштейн, Крутков и др., см. также
/[3]/) о том, что некоррелированной слу-
чайной функцией является не скорость,
а ускорение, т.е. сила. Точнее, не-
коррелированной предполагалась неупоря-
доченная сила, остающаяся после вычита-
ния гидродинамической силы, выражающей-
ся по формуле Стокса. Если, учитывая
гидродинамическое последействие, упоря-
доченную силу выражать формулой Ерусси-
не, то предположение о некоррелирован-
ности неупорядоченной силы приведет, в
частности, к результатам работы.
Физическим основанием предположения о
некоррелированности неупорядоченной
силы является малость её времени корре-
ляции по сравнению со временем релакса-
ции скорости для больших броуновских
частиц (большая частота ударов). Для
небольших частиц, когда время корреляции
сравнимо с временем релаксации, подоб-
ные теории не применимы. Но даже если
указанное предположение законно и тео-
рия правильно описывает процессы, про-
текающие в промежутки времени порядка
времени релаксации (и больше), то она
всё-таки является не пригодной для рас-
смотрения приращений скорости в течение
времён порядка времени корреляции не-
упорядоченной силы.

STANDARD TRANSLATION

In the first works on the theories of the
Brownian movement (see also #2) the values of
the velocity of a particle at various instants of
time were actually assumed to be statistically
independent. Accordingly, Einstein's formula
$M(x-x_0)^2= 2....(1)$ was applicable as well as the
Einstein-Fokker-Plank equation, which holds
true for Markov's processes. In reality, how-
ever, the correlation between the values of the
velocity is absent only at sufficiently large in-
tervals of time between the observed instants.
Therefore, formula (1) proves to be incorrect
for small intervals of time (of the order of mag-
nitude of correlation time for the velocity).

In order to formulate a more complete theory
which would be applicable for smaller intervals
of time, assumptions were made (Ornstein,
Kroutkou and others; see also #3) that the uncor-
related, random function is not the velocity, but
the acceleration, i.e., the force. More precise-
ly, it was assumed that the random force which
remains after the subtraction of the hydrodyna-
mic force, expressed by Stoke's formula, is un-
correlated. If by taking into account the hydro-
dynamic after-effect, the correlated force, is
to be expressed by Bousett's formula, then the
assumption of the uncorrelativity of the random
force will lead, in particular, to the results of
the work (perhaps he means to the satisfying
results?). The physical reason of the assump-
tion about the uncorrelativity of the random
force, is the shortness of time of its correlation
as compared to the relaxation time of the velo-
city of the large Brownian particles (high fre-
quency of collisions). For the small particles,
when the time of correlation approximates the
relaxation time, such theories are not applicable.
But even if the indicated assumption is legiti-
mate and the theory correctly describes the
process which takes place in the interval within
the relaxation time (and longer), the theory still
is not applicable for the observed increments of
velocity during the periods within the time of
correlation of the random force.

SIMULATED MECHANICAL TRANSLATION

In the first works on the theory of the Brownian movement (see also) the values of the velocity of the particle in the various moments of the time were assumed to be actually statistically independent. Accordingly, was applicable the formula of the Einstein and also the equation of the Einstein-Fokker-Plank, correct for the Markov's processes. In reality, however, the correlation between the values of the velocity is absent only at sufficiently large intervals of the time between the observed instants. Therefore, formula (1) proves to be incorrect for the small intervals of the time (within the time of the correlation for the velocity).

In order to create a more complete theory, applicable for the smaller intervals of the time, assumptions were made (Ornshtein, the Kroutkov, and others, see also) that the uncorrelated random function is not the velocity, and the acceleration, i.e., the force. More precisely, it was assumed that the random force, remaining after the subtraction of the hydrodynamic force, expressed by the formula of the Stokes is uncorrelated. If, by taking into account hydrodynamic after-effect, correlated force is to be expressed by the formula of the Boussinet, then the assumption about the random force will lead, in particular, to the results of the work. The physical reason of the assumption about the uncorrelativity of the random force is the shortness of its time of the correlation as compared with the time of the relaxation of the velocity for the large Brownian particles (large frequency of the collisions). For the small particles, when the time of the correlation is comparable with the time of the relaxation, such theories are not applicable. But even if the indicated assumption is legitimate and the theory correctly describes the process, taking place in the interval of the time within the time of the relaxation (and more), then it is, nevertheless, not applicable for the examination of the instants of the velocity during the period within the time of the correlation of the random force.

INSTRUCTIONS: <u>0</u> blank translation

("ending" means entire ending - not just final letter.)

1. Compare word with dictionary: If there is exact equivalence, translate. If there is multiple meaning, then this will be true for several consecutive words. In this case, choose the highest meaning common to all of the words. E.g., if there is a sequence of two words, the first having two meanings and the second three, then choose the second meaning for both.
2. If there is no exact equivalent, then remove as many letters from the end as is necessary to obtain a correspondence, and translate using the following rules. If there is no rule applicable to the ending, translate the word and ignore the ending.

RULES: The placement of "the". Place "the":

1. Before all nouns after a punctuation mark and before all adjectives when they begin a sentence.
2. Before nouns preceded by minor parts of speech and before adjectives also preceded by minor parts of speech except <u>не </u>.
3. After the verb, if the noun follows the verb or it is separated by one word.

Nouns preceded by adjectives:

1. If the adjective ending is ые , ых , их, и, then the noun is plural: otherwise sing.
2. If the word preceding the adjective is a noun, and if there is no punctuation mark between the first noun and the adjective, then place "of the" before the adjective.

Nouns preceded by pronouns:

1. Precede the pronoun by "of".

Nouns preceded by nouns:

1. If there is no punctuation mark between the nouns, then preface the second noun by "of the".

KOUTSOUDAS *from page 51*

Nouns preceded by punctuation:

1. If the noun ends in я , then hold translation until the verb is translated.   If the verb is plural, then the noun is plural, otherwise the noun is singular.

Nouns preceded by verbs:

1. If the word preceding the verb is not a noun, then invert the verb - noun word order.

Verbs preceded by nouns:

1. If the noun ends in y , then replace the "to" associated with the verb by "is to be".

Adjectives:

1. If the ending is ы , then precede the adjective by "are".

2. If the ending is   o , then precede the adjective by "is".

Verbs preceded by adjectives:

1. Preface the adjective by "is" and place at the end of the sentence; enclose the verb in "it --- that".