WORD

VOLUME 13 *December, 1957* NUMBER 3

AMBIGUITY OF SYNTACTIC FUNCTION RESOLVED BY LINEAR CONTEXT*

ANDREAS KOUTSOUDAS AND ASSYA HUMECKY

PROBLEM AND APPROACH

It should no longer be doubted that electronic computers will eventually be able to translate a scientific text from a foreign language into English, or into any other language. Recent progress in the fields of electronics and logical design has made it possible to solve by a digital computer both simple arithmetic questions and highly complex problems which were once thought too time-consuming for consideration. Engineers are now producing computers which have the speed and flexibility to carry out almost any set of unambiguous instructions.

The problem of translation, therefore, is not one of computer mechanics, but rather that of formulating unambiguous instructions. The preparation of such instructions necessitates analysis of (1) morphology, (2) syntax, and (3) multiple meaning, for both source language and target language. Translation by computer involves the formulation of a set of operational rules which would enable the computer to recognize a foreign word in the source language, to assign it its proper syntactic role within the target-language sentence, and, finally, to choose its proper target-language meaning for the given context.

This problem is here illustrated by a Russian-English example. In Russian, adjective stems (which may be participles) appear with suffix *-o* (or *-e* after palatalized consonants) functioning either as the «short form» of the neuter singular adjective or as

* This study was conducted at the University of Michigan under the supervision of Andreas Koutsoudas and with the research funds provided by the Engineering Research Institute.

an adverb. Depending on its function, such a construction requires different translation in English. For example :

Količestvo postojanno (adjective)='the quantity is constant'.

Količestvo menjaetsja postojanno (adverb)='the quantity changes constantly'.

Similarly adjective forms in *-ee* (or *-e* in certain morphophonemic circumstances) function as comparatives either of the adjective as a predicative complement or of the adverb as a modifier:

Zelezo slabee (adjective)='iron is weaker'.

Zelezo soprotivljaetsja slabee (adverb)='iron resists more weakly'.

Our present concern is thus with morphology and syntax insofar as the two meet under the area of homographs. In particular we are concerned with the problem of formulating such instructions as will enable an electronic computer (1) to identify the Russian *-o/-e/-ee* suffixed adverb, short-form adjective or participle in its respective function as either an adverbial modifier[1] or a predicative complement and (2) to supply its correct English equivalent.

The formulation of such instructions leads us to two considerations. The first is the classification of dictionary words into mutually exclusive classes. The second is devising a method by which the computer will recognize the need to insert or omit a certain group of words (such as the present tense of the verb 'to be', 'than', etc.) whenever these words are absent or redundant in Russian but must be supplied or omitted in the English translation.

It should be obvious that one of the important requirements for the Mechanical Translation of Russian is either a bilingual dictionary or a bilingual micro-glossary[2]. The micro-glossary would have to contain, along with the Russian words and their English equivalents, a set of distinct symbols for identifying the form-class of each particular word. The form-class identification within the dictionary becomes indispensable to the translation of meaning, singular or plural number, the present or past tense, etc. To be

[1] We shall use the term "adverb" only in the morphological sense and "adverbial modifier" in the syntactical (cf. Unbegaun, B. O., *Russian Grammar*, Oxford, 1957, pp. 289 and 293).

[2] The term "micro-glossary" was first used by V. A. Oswald to denote a dictionary composed out of a branch of a given scientific field — for example, Atomic Physics. See V. A. Oswald, "Microsemantics", mimeographed at the University of California at Los Angeles, 10 pages, June, 1952.

effective, the form-class division should be clear cut, with each class member belonging to one and only one form-class. Failure to provide for mutually exclusive form-classes would lead to ambiguity, first in regard to the correct choice of form-class, secondly with regard to a syntactic operation on the form-class. For example, if the word *zakonno* were listed in the form-class «adverb» as well as in the form-class «short-form adjective», then the computer would not know whether to choose the adverbial equivalent 'lawfully' or the adjectival equivalent 'lawful'. If the computer is led to treat this word as an adverbial modifier, then it is sufficient merely to supply its English equivalent. If, however, this word should be treated as a predicative complement, then it is also required that the English equivalent be prefaced by 'is' whenever the verb 'to be' is absent in the Russian context.

A further complication arises when the computer has to operate on a comparative adjective ending in *-e* or *-ee*. If any such word is identified as a predicative complement, that will mean its English equivalent must (1) be prefaced by 'is' or 'are' (under the above stated condition concerning the absence of the verb 'to be'), (2) be followed by 'than' whenever such provision is absent in the Russian syntax (genitive of comparison) but must be supplied in English, (3) ignore the translation of *čem* whenever it precedes the predicative complement and (4) be preceded by 'the'. Now in Russian, the distinction of form-classes is to a great degree dependent on suffixes[3]. Given a certain Russian word, one can usually identify it as a noun or a verb by merely checking its suffix. However, there is a considerable overlapping of suffixes, which sometimes makes it impossible to distinguish one form-class from another in this way. Such is the case with the short and comparative forms of the adjective in its adverbial or predicative function, involving the following suffixes: *-o/-e/-ee*.

In translating from scientific Russian, one finds that the English present tense of the verb 'to be' is most frequently the equivalent of the Russian words *n'et* or *'est* or of a dash, or is supplied where Russian uses a short adjective,[4] without any explicit copula. A method must therefore be devised for the electronic

[3] By "suffixes" we mean here both flectional and derivational morphemes, i. e., both "suffixes" and "endings".

[4] There are other possibilities for expressing the present tense of "to be" with which we are not concerned here, such as sentences of the type "*She* is a *student*", "it *is not* here, but there", etc.

computer to recognize and translate these four signals. For the words *n'el* and *'esl* the difficulty lies in choosing the proper meaning (*n'el*='is not', 'there is not', 'are not', 'there are not' and *'esl*='is', 'are', 'there is', 'there are', or 'to eat') for the given sentence or clause. Consequently, the problem is mainly that of multiple meaning rather than of syntax or morphology and need not concern us here. In cases of dashes, short and comparative forms of the adjective, recognition becomes a little more difficult; the computer has to recognize the fact that both the adjective and what immediately follows the dash function as predicative complements, and that it must supply therefore the proper form of the verb 'to be'[5].

To our knowledge, the only mention of this problem was made in K. E. Harper's report on the preliminary study of Russian for Mechanical Translation[6]. The solution offered was to treat the ambiguous Russian adverb as a short-form adjective, to preface both by «to be» when translating into English, and to let the reader make the proper choice. No mention was made of the dash.

The Experiment

A sample of 31,000 running words of Russian scientific text[7] was studied in an attempt to formulate rules for identifying and translating, by sheer mechanical means, the short-form participle or adjective[8] functioning as predicative complement or adverbial modifier.

The analysis of the ambiguous forms in question provided a few simple rules, which, in the majority of cases, prove sufficient to determine the syntactic function of these forms in a maximum environment of one word before and one word after the ambiguous form(s).

[5] Research in utilizing the dash as a physical form signaling a syntactic function has already begun in our laboratory but has not, as yet, been brought to completion.

[6] K. E. Harper, "A Preliminary Study of Russian," *Mechanical Translation of Languages*, The Technology Press of Mass. Institte of Technology and John Wiley and Sons, Inc., New York, 1955.

[7] Our sample was chosen from *Žhurnal Èksperimental'noj i Teoretičeskoj Fiziki*, Vol. 28, No 1, pp. 1-128, 1955.

[8] The short-form adjective in -o and the adverb -o as well as the comparative degree of both adjective and adverb, which are homographs, are treated by most grammarians as the same part of speech, namely adjective.

The investigation of the short-form Past Passive Participles ending in *-o* reveals a peculiarity requiring special attention: the adverbial form always ends in *-nno*, while the predicative form has a single *n(-n-)* before the flectional suffix. It must be borne in mind that in Russian there are also adjectives bearing either of these suffixes (e. g., *zakonno* and *ravno*) without, however, the predicative-adverbial dichotomy being involved. The first step then was to work out a method of differentiating between these overlapping suffixes, and the following solution was found: all adjectives and participles would be grouped together as «Adjective Class Members»[9]. In the «memory» (electronic dictionary) of the machine, the adjectival stems would include the *n* or *nn*, as the case may be (e. g. *zakonn-*», «*ravn-*»), while the participial stems would be stripped of these suffixes (e. g. *svjaza-*). Consequently, *-o/-no/-nno* would be treated as three different suffixes. This will enable the computer automatically to identify the syntactic function of a word merely by checking its suffix: if *-no*, predicate; if *-nno*, adverbial modifier.

Having eliminated the problem of the Past Passive Participle, the next step was the analysis of *-o/-e/-ee* forms of the remaining Adjective Class Members.

The material was divided into three groups respectively, containing: (1) a single *-o/-e/-ee* form functioning as a predicative complement, (2) a single *-o/-e/-ee* form functioning as adverbial modifier and (3) any two of these forms occurring consecutively. Furthermore, these groups were limited, whenever possible, to three units: the central word—the form(s) under question *(-o/-e/-ee)*—and the adjacent ones to the left and to the right from the central. At the same time it was noticed that certain adjectives, regardless of their position, tended to appear in only one capacity—either adverbial modifier or predicative complement—while still others appeared as prepositions. All these instances were listed separately.

The analysis led to the discovery of the following preliminary rules :

I. *Predicative Complement*

Left	Central	Right	Number of Occurrences
1. not verb	*-e/-ee*	Adjective Class Member (ACM)	5

[9] We shall use the term "adjective" in its regular grammatical meaning and "Adjective Class Members" (ACM) to include the adjectives and participles.

Example: metallax krajne *zatrudnitel'no*
in metals *is* extremely *difficult*
naxoždenie ešče bolee *zatrudnitel'no*
finding *is* still more *difficult*

	Left	Central	Right	Number of Occurrences
2. *a)*	not verb	*-e/-ee*	čem	14
	Example: gde temperatura *niže,* čem where the temperature *is lower* than			
b)	«čem/«čto»	*-e/-ee*	not verb	
	Example: , *čem bol' še* X (a formula) the *greater is* X , čto *složnee* v slučajax , which *is* more *complicated* in cases			
3.	not verb, nor «esli»	*-o/-e/-ee*	infinitive	45
	Example: , *legko* polučit' , it *is easy* to obtain Poétomu *estestvennee* vybrat' Therefore it *is more natural* to choose			
4.	formula, or a noun, or, «éto» or «čto», or «to»	*-o/-e/-ee*	preposition or a pronoun	18
	Example: pole *perpendikuljarno* k the field *is perpendicular* to éto *zakonno* dlja this *is legitimate* for , to *niže* ee , now it *is lower* than it			
5.	formula or noun	*-o/-e/-ee*	noun	18
	Example: Javlenie *analogično* éffektu (the) phenomenon *is analogous* to the effect 9 *sil'nee* énergii 9 *is stronger* than the energy			
6.	comma	*-o/-e/-ee*	period or bracket	3
	Example: , *nedopustimo)* , *is* not *admissible)*			
7.	not verb or ACM	*-o/-e*	formula (on the same or the next line) or a colon	17

Example: X *niže* X
X *is lower* than X
X *ravno* X
X *is equal* to X

	Left	Central	Right	Number of Occurrences
8.	formula or noun	*-o*	period or comma or bracket	14

Example: X *postojanno.*
X *is constant*
rezonansa *izvestno,*
of the resonance *is known,*

9.	not verb	*-o*	comma «čto»/«čem»	25

Example: *Xarakterno,* čto
It *is characteristic,* that
bolee *zatrudnitel'no,* čem
is more *difficult,* than

10.	not verb	*-o*	étoj/étomu/étim	2

Example: takže *proporcional'*no étoj
is also *proportionate* to this

11.	«kotoroe»	*-o*	ACM or noun	2

Example: kotoroe *proporcional'no* raznosti
which *is proportionate* to the difference

12.	«kak»	*-o*	verb infinitive or any non-verb	22

Example: Kak *legko* pokazat'
As it *is easy* to show
kak *izvestno,*
as it *is known,*

Total: 185

II. *Adverbial Modifiers*

	Left	Central	Right	Number of Occurrences
1.	Verb other than infinitive (could be followed by «tem»)	*-o/-e/-ee*	irrelevant	38

Example: ocenivalos' *vizual'no* po
was *visually* determined by

otklonjaetsja tem *sil'nee*, čem
deviates the *stronger*, the

	Left	Central	Right	Number of Occurrences
2.	period or comma («a» or «kak i» could follow)	*-o/-e/-ee*	comma not followed by «čto» or «čem»	33
	Example: , *estestvenno*, , *naturally*, *Dalee*, kak *Further*, as , a *vozmožno*, i , and *possibly*, also			
3.	not «kak»	*-o/-e/-ee*	verb other than infinitive	78
	Example: , *suščestvenno* menjaet , *substantially* alters			
4.	ACM	*-o/-e/-ee*	irrelevant	34
	Example: raspoložennym *perpendikuljarno* k placed *perpendicularly* to			
5.	«esli» or auxiliary verbs or words**	*-o*	infinitive	28
	Example: Esli *otdel'no* najti If we find *separately* *budet slabo* zaviset' will *slightly* depend *dolžno sil'no* zaviset' must *strongly* depend			
6.	irrelevant	*-o*	ACM	94
	Example: v *sravnitel'no* redkix in *comparatively* rare			
7.	dash	*-o*	irrelevant	5
	Example: — *približenno.* — *approximately.*			
			Total:	310

** These include words like *možno*, *nel'zja*, *možet byt'*, *dolžny*, *neobxodimo*, as well as *vozmožnost'*, *neobxodimost'*, etc.

III. Sequence of two *-o* or one *-o*, one *-e/-ee*

	Left	Central	Right	Number of Occurrences
1.	not verb or ACM	*-o/-e/-ee*	not verb or ACM	17

Example: metallax *krajne zatrudnitel'no* iz-za
in metals is *extremely difficult* because of
sorta *dostatočno malo.*
sort is *sufficiently small.*

(rule: first of the two is adverbial modifier, second—predicative complement)

2.	irrelevant	*-o/-e/-ee*	verb or ACM	14

Example: *dostatočno bystro* ubyvaet
sufficiently quickly diminishes
nedostatočno experimental'no obosnovan.
insufficiently experimentally grounded.

(or vice versa: two *-o* preceded by verb or ACM rule: both are adverbial modifiers.)

3. a) *Bolee* or *menee* are always adverbial modifiers; apply rules III,1 and III,2 to identify the remaining form. 5

Example: see last example of I,1.

b) *Bol'še* or *men'še* preceded by the verb 'to be' (any tense) are always predicative complements; the other *-o/-e/-ee* forms automatically become adverbial modifiers.

Example: makroskopičeskogo *značitel'no men'še* of the macroscopical *is considerably smaller*

Total: 36

These rules are applicable to almost all cases, but they were considerably redundant and had to be reduced to one set of yes-or-no choices. We decided to classify first all homographs in our «Adjective Class», giving to each two separate translations: (1) adverbial and (2) adjectival. For example, the word *zakonno* will be listed in the «Adjective Class» and will have translations (1) 'lawfully' and (2) 'lawful'. We then formulated a set of final

rules based on those of lists I and II to specify the conditions under which the adverbial translation would be chosen. It is assumed that the word in question has already been identified as a member of the «Adjective Class». This word is to be considered together with the preceding and following unit. A unit is the word or punctuation mark occurring immediately before (or after), not counting: (1) parentheses or words in parentheses, (2) all non-*-o/-e* adverbs, (3) 'to' preceded by a comma, (4) a formula preceded by a preposition, (5) the words *bolee* and *menee* (which are always adverbs), and (6) the words *tem, a, kak i, libo, ne, že*. This set of rules omits certain instances of infrequent occurrence, and results in correct translation 98.5% of the time (676 out of 686 cases).

1. Choose the adverbial translation of the following words, if they are preceded or followed by a verb other than 'to be': *ravno, spravedlivo, bol'še, men'še.*

Choose the adverbial translation if:

2. a) the word ends in *-o/-e/-ee* and is preceded by an Adjective Class Member other than *kotoroe* or *ėto*. Example: see that of II,4.

   b) the word ends in -o and is followed by an Adjective Class Member. Example: see that of II,6.

3. a) the word ends in *-o/-e/-ee* and is preceded or followed by a verb (not auxiliary) other than an infinitive.

   Example: ocenivalos' *vizual'no* po
   was *visually* determined by
   suščestvenno menjaet
   substantially changes

   b) the word ends in *-o/-e/-ee* and is preceded by *esli*, auxiliary verb(s), *vozmožnost'*, *neobxodimost'*, etc. and followed by an infinitive. Example: see II,5.

4. the word ends in *-o/-e/-ee* and is preceded by a dash, period, or comma and is followed by a comma which in turn is not followed by «čem», «čto», or «tak kak». Example: see II,2 and 7.

5. If any two consecutive words end in *-o/-e/-ee*, then choose the adverbial translation for both if the above rules (2 or 3) apply. Choose the adverbial translation for the first and the predicative one for the second if the above rules (2 and 3) do not apply Example: see III.

6. All the instances to which the above rules do not apply automatically become predicates and the computer will preface them by 'is'.

Once a word is established as a predicative complement, some additional rules determine the placement of pronouns, articles, and the verb 'to be'. These rules are as follows:

a) In a sequence *čem* — *-e/-ee* put 'is' after *-e/-ee* and 'the' before it; do not translate *čem*.

Example: *čem* bol'še X
the greater *is* X

b) In a sequence not *tem* — *-e/-ee* — *čem*, put 'is' before *-e/-ee;* translate *čem*.

Example: gde temperatura niže, *čem*
where the temperature *is* lower, *than*

c) In a sequence *tem* — *-e/-ee* — *čem*, put 'the' before and after *-e/-ee;* do not translate *čem* and *tem*.

Example: proisxodit *tem* bystree, *čem*
takes place *the* sooner, *the*

d) In a sequence 'to be' (any tense) — *-e/-ee* — ACM, put 'than the' after *-e/-ee*.

Example: budet ne *niže* pervogo (porjadka)
will not be *lower than the* first (order)

e) In a sequence *bylo/budet* — *-o/-e/-ee* — formula or noun, put 'than' after *-o/-e/-ee* and choose the adjective translation but do not preface it by 'is'.

Example: *bylo* ne *niže* X
was not *lower than* X

f) In sequences *kak* — *-o*, and «non-verb» or *esli* — *-o* — infinitive verb, preface 'is' by 'it'.

Example: kak *izvestno*
as *it* is known

Netrudno pokazat'
It is not *difficult* to show

We also discovered a list of words which, in our sample, always had a single function. These words and their frequencies, are the following.

1. Always adverbial modifiers: *otdel'no* (4), *neposredstvenno* (4)

nezavisimo (19), *kačestvenno* (6), *obyčno* (17), *okončatel'no* (8), *sootvetstvenno* (18), *primerno* (7), *ranee* (6), *dalee* (5), *podobno tomu kak* (1), *bolee* (3), *menee* (3).

2. Always predicative complement: *veliko* (4).
3. Always prepositions: *soglasno* (17), *otnositel'no* (28).

All these words will be listed in their complete form under their respective form-class (Adverbs, Verbs, and Prepositions).

Three additional rules were constructed, on the basis of regularities observed, for the remaining 10 instances (or 1.5%) not accounted for by our general set of rules. They are the following:

1. If among the three words which precede the *-o/-e/-ee* word none is a noun, or a personal pronoun, or an adjective in the nominative case (in case of an *-o* word — none is neuter singular nominative), the *-o/-e/-ee* word is an adverb (8 cases).
2. If an *-o* word is preceded by a period and followed by a noun and a comma, it is an adverb (1 case).
3. If the word *malo* is followed by an adjective and the word *mesta*, translate it: 'there is little... space' (1 case).

Thus, there remain no exceptions to our rules among 686 cases in the 31,000 word sample examined so far. It is anticipated that the inclusion of the last three rules or the expansion of the entire set of rules may be required when larger samples of text will be analyzed (a new sample text of 42,000 words is now being analyzed). On the other hand, rules such as (3) above, which apply to extremely rare cases, might be omitted from the final scheme on the grounds that 99% or 99.9% accuracy is sufficient.

University of Michigan.