2

"1965 International Conference on Computational Linguistics"

SUBCLASSIFICATION OF PARTS OF SPEECH IN RUSSIAN: VERBS*

A. Andreyewsky

International Business Machines Corporation
Thomas J. Watson Research Center
P. O. Box 218
Yorktown Heights, New York, 10598

*This work was partly sponsored by the Information Processing Laboratory, Rome Air Development Center, United States Air Force, under Contract AF 30(602)-3301

Abstract

In a trial study, about 500 Russian verbs were coded using 44 potential classificatory criteria. Through sorting and the introduction of a metric, numerous groupings were obtained. Initial results suggest that, with proper refinements, the approach described can provide useful information that may be employed in syntactic analysis and certain information retrieval applications.

0.0 Introduction

As part of a broader effort to extend the existing traditional part-of-speech classification in modern Russian, this study of verbs is oriented toward developing an improved basis for syntactic analysis. Moreover, it is hoped that the refinements introduced will be of interest in content analysis. To this end, an extensive set of potential classificatory criteria has been selected, in the hope that eventually this categorization can be optimized and extended to other parts of speech.

1.0 The Experiment

The 514 verbs analyzed came from two sources: (a) a randomized sample of 370 entries (1) and (b) a list of the most frequently used Russian verbs (2) from which the first 144 entries were selected.

The classificatory criteria, subdivided into two groups, are discussed in Section 1.1 below. Generally, each verb was taken in a particular meaning (stirat', for instance, as "to erase" and not as "to launder") and English equivalents used solely for purposes of identification. At the same time, for reasons of convenience, provisions were made in coding to allow for coexisting alternatives. Thus, for properties A and B there can be four possibilities which are represented by the following numerical codes: 1 - "A", 2 - "B", 3 - "AB", 0 - "neither applies".

After the verbs and appropriate codes were punched on cards, verbs with identical codes were compared. To obtain additional clustering, a program, written by R. F. Hubbard for the IBM 7040, compared the code vector of each card against those of the rest of the sample. The distance between any two entries was calculated by taking the square root of the sum of the squares of distance between corresponding positions in their code vectors as defined by the following table:

$0 \rightarrow 0 = 0 \quad 0 \rightarrow 2 = 4 \quad 1 \rightarrow 1 = 0 \quad 1 \rightarrow 3 = 1 \quad 2 \rightarrow 3 = 1$
$0 \rightarrow 1 = 4 \quad 0 \rightarrow 3 = 6 \quad 1 \rightarrow 2 = 2 \quad 2 \rightarrow 2 = 0 \quad 3 \rightarrow 3 = 0$

1.1 Tests

Since one of the main objectives of this study has been to establish the relevance of various classificatory criteria, these were tested in two groups as described below. The selection of criteria, based on studies of existing grammars of Russian, was directed toward discovering solutions for problems arising or likely to arise in machine-assisted syntactic analysis.

1.1.1 Test I

In this test, the verbs were coded according to their ability to combine with selected prepositional phrases, certain adverbs, and the chto-introduced object clauses. Most of the examples are derived from the discussion of slovosochetaniye (grammatically bound word group) problem in the Academy Grammar (3). While the English meanings supplied do reflect certain semantic differences the main objective has been to test not only the ability of a given verb to co-occur with certain types of phrases (examples are used solely for illustration) or classes of adverbs but to trace what effect the verb has on their syntactic function, if any.

1.1.1.1 Classificatory Criteria

1) ... do menya
   (A) before me
   (B) as far as me

2) ... do rassveta
   (A) before dawn
   (B) until dawn

3) ... iz-za stola
   (A) because of ...
   (B) from behind the table

4) ... k mitingu
   (A) for the ...
   (B) to the ... meeting

5) ... k nam
   (A) to us
   (B) toward us

6) ... za obedom
   (A) after (to get)...
   (B) during dinner

7) ... u Ziny
   (A) at Zina's
   (B) from Zina

8) ... pod kapustu
   (A) for cabbage
   (B) under cabbage

9) ... za stol
   (A) at the table
   (B) behind the table

10) ... za brata
(A) in brother's place
(B) for brother's sake

11) ... po oshibke
(A) a mistake apiece
(B) by mistake

12) ... yashchik iz-pod uglya
(A) coal crate
(B) crate from under the coal

13) ... o stol*
against the table

14) ... po vodu*
to get water

15) ... chto napishet*
that + (subject) + will write

16) ... nadvoe*
in two (as in cutting)

17) ... ochen'*
very much

18) ... o sestre*
about the sister

1.1.1.2 Results of Sorting

Sorting revealed some of the following groupings with identical codes:

A1 zakipet'
(to boil)
prosolit'sya
(to turn salty)

A2 vzdrognut'
(to flinch)
ustavat'
(to become tired)
ustat'
(to become tired)
izzyabnut'
(to become chilled)

A3 sovrat'
(to tell a lie)
soobrazit'
(to grasp)
dogadat'sya
(to surmise)
nameknut'
(to hint)

A4 raznuzdat'
(to let become undisciplined)
vospitat'
(to educate)

A5 vychest'
(to subtract)
izderzhat'
(to spend)

A6 otrubit'
(to chop off)
vskryt'
(to open up)

A7 nabryzgat'
(to sprinkle on)
rasprostranit'
(to spread)

A8 vbezhat'
(to run in)
yavit'sya
(to appear)

A9 podumat'
(to think)
bredit'
(to rave)
okhat'
(to moan)

A10 gordit'sya
(to be proud)
veselit'sya
(to enjoy self)
voskhishchatsya
(to admire)

A11 verit'
(to believe)
toskovat'
(to be sad, to pine)

* Only test ability to combine in the meaning indicated.

grustit'
(to be sad, to yearn)
skuchat'
(to be bored)
fantazirovat'
(to dream)

A12 volnovat'sya
(to worry)
opasat'sya
(to be afraid)

A13 zapryatyvat'
(to hide)
vovlekat'
(to draw in)

A14 berech'
(to save)
poberech'
(to save)
uderzhat'
(to withhold)

A15 vzgromozdit'sya
(to perch self
uglubit'sya
(to go deep into)
rassazhivat'
(to seat)

A16 nastroit'
(to incite)
bespokoit'
(to disturb)
obizhat'
(to offend)
proklinat'
(to damn)
portit'
(to spoil, ruin)

A17 bakhvalit'sya
(to brag)
likovat'
(to rejoice)

A18 razrykhlyat'
(to loosen)
razdrobit'
(to pulverize)

A19 besedovat'
(to converse)
soveshchat'sya
(to confer)

A20 razodrat'
(to tear)
rasshibit'
(to break, bust)

A21 morosit'
(to drizzle)
nakrapyvat'
(to sprinkle)
poroshit'
(to snow)

A22 farshirovat'
(to stuff)
sintezirovat'
(to synthesize)

A23 klassifitsirovat'
(to classify)
razbivat'
(to break)

A24 begat'
(to run)
prikhodit'
(to come)

A25 nesti
(to carry)
vezti
(to cart)

A26 volochit'
(to drag)
tashchit'
(to pull)

A27 doyti
(to reach, (walking))
doletet'
(to reach (flying))

1.1.1.3 Results of the Introduction of the Metric

On the basis of preliminary results, the maximum distance considered was set at 10. Given this arbitrary limitation, the metric produced various groupings. The majority of them contained some "noise" - i.e., apparently incorrect entries were brought together or several distince groupings turned out insufficiently differentiated. Partly responsible for this are: the method employed, the distances selected, and the occasional errors that crept in during the analysis and subsequent processing. These factors are discussed in greater detail below (1.1.1.4).

Some of the more interesting outcomes were as follows:

A28 Groups A11 (verit', toskovat', grustit', skuchat', and fantazirovat'), A17 (bakhvalit'sya and likovat'), A12 (volnovat'sya and opasat'sya), and the verb bespokoit'sya (to worry)

A29 Group A16 (nastroit', bespokoit', obizhat', proklinat', portit') and the verb nenavidet' (to hate).

A30 Group A10 (voskhishchyat'sya, veselit'sya, gordit'sya) and verbs vozmutit'sya (to become disgusted) and boyat'sya (to be afraid).

A31 Group A8 (yavit'sya, vbezhat'), the following verbs: vernut'sya (to return), prikhodit' (A24), begat' (A24), vyyti (to step out), podyezzhat' (to drive up), yezdit' (to ride), vyekhat' (to go away), kinut'say (to lunge), vypolzti (to crawl out), doletet' (A27), and doyti (A27).

A32 garantirovat' (to guarantee), pokazyvat' (to show), demonstrirovat' (to demonstrate)

A33 sovrat' (to lie), poverit' (to believe), uverit' (to assure)

A34 znat' (to know), ozhidat' (to expect), videt' (to see).

A35 nagryanut' (to come unexpectedly), zaekhat' (to stop by), probezhat'sya (to run), otstupit' (to retreat).

1.1.1.4 Comments

The problems stemming from the application of the metric (th "numbers game") mentioned in 1.1.1.3 reflect a characteristic of statistical inference jocularly compared by an anonymous author to a bikini bathing suit: being sufficiently suggestive, but not revealing. In this regard, alternative approaches have been considered and will be tried in the near future. As it turned out in practice, however, the metric did provide useful insights which can point the way toward developing a more powerful set of classificatory criteria. This, in turn, can foster increased reliance on simple sorting procedures based on proper ranking and grouping of the criteria themselves.

While not unexpectedly, the verbs of motion in the broad sense of the term came out more clearly in the classification than did any other groups, interesting subclasses of abstract verbs, exhibiting unexpected shades of valuation also emerged.

1.1.2 Test II

In contrast to Test I, this test placed a relatively lesser emphasis on syntagmatic relationships and stressed a mixture of formal and semantic properties. On the whole, except where noted, the two tests were developed independently of one another. While Test I was based on materials derived from the Academy Grammar of Russian (3), Test II benefited from experience gained in dealing with the problems encountered in machine translation output and from studies conducted preparatory to launching syntactic analysis.

1.1.2.1 Classificatory Criteria

In view of the extensive nature of this test, the description of various criteria used is given here in abbreviated notation.

1) (A) imperfective
   (B) perfective
2) verb (1/3) or "verboid" (2/0); "concrete" (1/0) or "abstract" (2/3): when "yes" answer is possible under 1.1.1.1.17.
3) is given form
   (A) reflexive
   (B) non-reflexive
4) generally:
   (A) non-reflexive
   (B) reflexive
5) when reflexive, meaning:
   (A) active
   (B) passive
6) participial forms:
   (A) active
   (B) passive
7) passive participle:
   (A) past
   (B) present
8) gerundial forms:
   (A) present
   (B) past
9) action (gerund):
   (A) parallel
   (B) sequential
10) deverbal nouns:
    (A) in -enie, -ka
    (B) other forms
11) deverbal nouns:
    (A) concrete
    (B) abstract
12) verb used:
    (A) personally
    (B) impersonally
13) verb function:
    (A) link, auxiliary
    (B) other
14) meaning affected by
    (A) governed infinitive
    (B) object(s)
15) subject preference:
    (A) inanimate
    (B) animate
16) verb governs:
    (A) infinitives
    (B) objects
17) object preference:
    (A) animate
    (B) inanimate
18) (A) motion verb (broad sense)
    (B) action perceived
19) verb describes:
    (A) action
    (B) state
20) (A) beginning
    (B) end of action

21) verb is one of:
(A) being
(B) becoming

22. action described:
(A) outward-
(B) inward-directed

23) action directed:
(A) downward (away)
(B) upward (toward)

24) action in respect to object:
(A) contacts
(B) permeates

25) reference to:
(A) duration
(B) intensity

26) action produces:
(A) decrease
(B) increase

27) action describes:
(A) gain
(B) loss

1. 1. 2. 2 Results of Sorting

The following groupings had identical codes:

B1 skuchat' (A11) (to be bored)
toskovat' (A11) (to be sad)

B2 morosit' (A21) (to drizzle)
poroshit' (A21) (to snow)

B3 nakrapyvat' (A21) (to sprinkle)
mertsat' (to twinkle)

B4 izderzhat' (to expend)
istratit' (to spend)

B5 nosit' (to carry)
tashchit' (to pull)
volochit' (to drag)

B6 podshivat' (to attach)
navyazyvat' (to tie on)
skladyvat' (to put together)

B7 ozhivit' (to vivfy)
uverit' (to assure)

1. 1. 2. 3 Results of the Introduction of the Metric

Comments made in 1. 1. 1. 3 above, apply. Because of a greater number of classificatory criteria the results of introducing the metric were more important in this test. Numbers in parentheses preceding each verb indicate distances from the first verb in the group.

B8 pridavit' (to squeeze)
(1) prishchemit' (to pinch)

B9 vosstat' (to riot)
(1) vystupit' (to appear)

B10 prichesat' (to comb)
(1) zaputat' (to tangle)

B11 vbezhat'
(to run in)
(1) vypolzti
(to crawl out)

B12 napevat'
(to hum)
(2) veshchat'
(to speak with authority)

B13 temnet'
(to grow dark)
(2) teplet'
(to grow warm)

B14 vyrabotat'
(to develop)
(3) vyuchit'
(to learn)

B15 khmurit'sya
(to frown)
(3) tumanit'sya
(to grow gloomy)

B16 razbushevat'sya
(to start raging)
(4) uchastit'sya
(to become more frequent)

B17 otkryt'
(to open)
(5) ubavit'
(to decrease)

B18 vynesti
(to carry out)
(5) vypustit'
(to let out)

B19 zheltet'
(to turn yellow)
(5) umirat'
(to die)

B20 terrorizirovat'
(to terrorize)
(5) khvalit'
(to praise)

B21 viset'
(to hang)
(6) lezhat'
(to lie)

B22 podognat'
(to drive up)
(6) navestit'
(to visit)

B23 nestis'
(to dash)
(7) bezhat'
(to run)

B24 vydelit'
(to single out)
(7) vypisat'
(to write out)

B25 potusknet'
(to dull)
(7) zatverdet'
(to harden)

B26 prikrepit'
(to fasten)
(8) nav'yuchit'
(to pack on)

B27 vozvratit'
(to return)
(8) dopolnit'
(to augment)

B28 vvesti
(to introduce)
(9) dobavit'
(to add)

In addition to shorter groups described above, longer groupings were observed. Thus, otdokhnut' (to rest) (8) utikhnut' (to quiet down), and (10) ugasnut' (to become extinguished) or nabryzgat' (to sprinkle on), (2) nakinut' (to throw on), (3) vzvalit' (to pile on), and (4) nastrocit' (to sew on) are some of the examples.

In other cases, apparently incongruous groups like the following: strekotat' (to chirr), (1) moshennicat' (to swindle), (5) fokusnichat' (to juggle), (5) nakrapyvat' (to sprinkle), (5) mertsat' (to twinkle) (6) zvenet' (to ring) emerged. However, upon closer examination it became apparent that nakrapyvat', mertsat', and zvenet' fall in a group clearly distinguishable from the one containing the other verbs. Further, fokusnichat' and zvenet' showed sufficient distance within respective groups suggesting at least four different basic groups in all.

1.1.2.4 Comments

Aside from the problems traceable to statistics, the sets of criteria selected for Test II are more open to debate than those found in Test I. However, correlations between both tests indicate that some of the criteria are relevant and that others are, at least, redundant. As observed from minor differences in two versions of coding of nine verbs introduced six months apart, the results of Test II are less reliable.

1.1.3 Comparison of Test I and Test II

As noted in 1.1.2 above, the two tests differ in the base from which they were derived. Accordingly, the results obtaining from Test I are both intuitively and actually more reliable. Yet, as suggested in 1.1.1.4, to the extent that the results of the application of the metric tend to supplement sorting, the results of Test II tend to back up many of the findings of Test I.

Given a small sample, it is difficult to make any generalizations. At the same time, the evidence emerging so far suggests some subtle differences in the two tests. Basically, in both cases the results of the metric application show little or no discrimination between antonyms. However, the groupings resulting from Test II tend to be, if at all, held together by similarity of content. the results of Test I, in contrast, have a peculiar sort of outward, formal similarity in the manifestation of processes described by the verbs in question.

2.0 The Outlook

In the months ahead, it is hoped that the small corpus can be increased and the time required to code each entry reduced to reasonable proportions. While in many respects the results of both tests are self-proving, rigorous evaluation criteria will have to be formulated in detail.

As far as potential application of the results obtained is concerned, especially the information derivable from Test I could be immediately put to use to improve (together with classification of nouns currently in progress) the translation of verb-governed prepositional phrases. It is likely that this syntagmatic patterning will extend to larger structures dominated by the verb. Further, if the apparent trends persist, some framework of semantic classification can be anticipated. To what extent this will be possible to accomplish by computers alone and the degree to which such a classification

will satisfy the needs of computer processing remains to be established. While it can be argued that any classification is likely to produce some classes, we take solace in the fact that the methodology employed even in such classics as Roget's Thesaurus remains unknown to this day.

Sources

1. This sample was selected from the Daum and Schenck Dictionary in another connection and was generally random in its intent more than its methodology.

2. A. K. Demidova, O. G. Motovilova, G. D. Shevchenko, E. P. Chaplygin, Naiboleye upotrebitel'nyye glagoly sovremennogo russkogo yazyka (The Most Frequently Used Verbs in Modern Russian), Moscow, USSR Academy of Sciences Publishing House, 1963.

3. V. V. Vinogradov, ed., Grammatika russkogo yazyra (Grammar of the Russian Language) Moscow, USSR Academy of Sciences Publishing House, 1960, Vol. II, Part I, pp. 113-230.