Computer Semantics of Russian

V.A. TUZOV
Department of Programming Technology
Saint Petersburg State University
Bibliotechnaya pl. 2
Saint Petersburg 198904
RUSSIA

Abstract: The brief summary of solving the problem of the computer semantic analysis of Russian texts is presented here. The essence of the analysis is the translation from Russian into the formal semantic language. The semantic dictionary which is the computer analogue of a learner's dictionary (replete with explanation) is used for translation. The dictionary contains more than a hundred thousands lexical units. Its basis consists of nearly twenty thousands basic concepts (a language thesaurus) divided into 1450 classes and a small collection of about two dozens basic functions.

Key-Words: Natural Language, Semantic Analysis, Formal Semantic Language

1 Introduction

There is a great deal of research connecting with the processing natural languages. Most of it has been based on models introduced in 70^s and 80^s. The most popular approaches to use in current applications include a) vector space model, b) boolean model, c) probabilistic model. These utilize statistical methods. Their comprehensive description can be found in [1, 2]. These approaches are out of favor: Presently, the standard "ad-hoc" task is not the major task presented at the important conferences on Information Retrieval such as TREC and NTCIR Workshop [3, 4]. The retrieval efficiency of the presently used systems cannot be significantly improved. The alternative functional model to a language formalization applied to the Russian language was proposed in 1979 by Melchuk I.A. [5]. It was simple and elegant. But it soon became clear that the model suggested in this book is inadequate to the Russian language. The absolutely right linguistic insight of Melchuk I.A. was in obvious contradiction with the formalism he used. It was necessary to solve this problem. And this was done, taking about 20 years to build the working model of the Russian language. A number of different pieces of research in the area of linguistics and beyond, affected the development of this model. It's impossible even to name all of them here. We cite only several studies which have not staled with the passage of time [6,7,8,9].

The developed model includes:

1) a language thesaurus – about twenty thousand basic concepts which are assigned to have no internal semantic structure;
2) a language basis – relations which can exist between entities and axioms of relation logic represent the allowed transformation rules for them;
3) a tree-based classifier includes about 1450 classes of word meanings, which provides the base for description of language lexis; each meaning corresponds to some node in this tree of classes and can have an arbitrary number of relations with other nodes;
4) a semantic dictionary that describes about 117 thousand lexemes (single words and phrases) in the semantic language. The dictionary describes both the internal structure of the word meaning and the information for calculating possible relations including restrictions for different kinds of relations;
5) a technique of semantic analysis which defines the algorithm of the text transformation into the semantic representation.

Here is a short example of using the developed semantic language for representing the meaning of words.

The word *адресовать [to address]* in the explanatory dictionary of Ozshegov S.I. is defined as *послать по какому-нибудь адресу [to send to some address]* though it is desirable to define it more precisely, such as *выполнить действие*

перемещения чего-либо куда-либо используя чей-то адрес [to execute an action to move something somewhere using someone's address]. The formula of this statement in the developed semantic language looks like follows:

N%~АДРЕС$12/0171(PerfCaus(Uzor(!Им,АДРЕС$12/0171(!Дат), #), Mov(!Вин, !Откуда, !Через, !Дат\!кДат\!Куда)))

The formula describes the meaning of the word *адресовать* and contains the subsequent information:

- the word belongs to semantic class адрес (АДРЕС$12/0171).
- the immediate interpretation: *to execute an action* (PerfCaus), *using*(Uzor) *address* (АДРЕС$12/0171) *to move*(Mov) *something*(!Вин) *somewhere*(!Дат\!кДат\!Куда).
- the word *адресовать* [to address] can be combined with a noun in nominative (!Им) in the role of actor, in dative (!Дат) in the role of address owner(АДРЕС$12/0171(!Дат)), with the noun in accusative denoting the object of movement (the first argument of Mov) and with the adverbial modifier of place including the compound (!Откуда).

The practical significance of the translation from Russian into the formal semantic language is that it allows putting a content of the text into knowledge bases. And it becomes possible then to use different formal methods to manipulate this content.

The presented technique of semantic analysis allows increasing the quality of many different intelligent systems, especially the following kinds of systems:

1) Intelligent text analysis systems (search engines, summarization systems, document filtering systems, knowledge-management systems)
2) High-quality machine translation systems
3) Speech recognition systems
4) Expert systems and computer-aided learning systems with natural language capabilities
5) Virtual reality systems and interactive films with capability of communication in natural language

The semantic dictionary was successfully used to develop search engines (http://www.soft.velton.net.ua/langs_download.html), text classification systems [13,18], summarization systems [14], and question-answering systems [15]. It was also applied to the automatic updating task of knowledge bases [16,17].

This paper contains a very brief overview of the proposed model. Its detailed description can be found in [10,11,12].

2 Functional language nature

The principal difference between the approach proposed here and all existing approaches to language formalization is that it uses a functional representation of the sentence meaning. From this point of view each word of the Russian language is the name of a function *f(x1,…,xn)* connected with this word and called its *semantics*. The word obtains each of its particular meanings only after the substitution of the particular arguments. The meaning of the word is calculated in the process of the function *f* execution. The sentence represents a single complete superposition of word functions. The meaning of the sentence is calculated in the process of the superposition's construction and execution.

The semantic analyzer performs two basic operations during the superposition's construction. The first of them makes the choice of proper word meaning (formal interpretation). And the second one combines chosen meanings to the meaningful subconstructions, i.e. to the constructions, which have their own independent semantic interpretation.

The requirement of an independent subconstrucions' interpretation results in the fact, that the combining of words can't be reduced to their simple composition. Functional word interaction occurs while combining, and a new meaning is calculated as result, in particular a new semantic-grammatical type of construction is built. For example, when the preposition "В" interacts with the phrase *"прошлом году"* the type "когда" is produced. In a case where the interaction of the word *"лесу"* with the aforementioned phrase the type "где" is selected, etc. The specific meaning is an empty meaning (or nonsense), which results in a semantic combination break. An interaction of a noun and an adjective produces a grammatical type of a noun; however, the phrase semantics may differ considerably from the noun semantic. For instance, such is the case in the phrase *"Белая ворона" [maverick]*. It concerns not only the adjective and the noun, but other words that comprise the sentence.

The frequently used linguistic notion of "valence" should be literally understood in the chemical sense: oxygen and hydrogen realizing their

Table 1. Semantic description of the word *"ГОРЯ" [mistery]*

Горя
ГОРЕ { Сущ Неодуш $150023~@ОНО$5@Род } $150023(!Род)
ГОРЕТЬ {Глаг} Caus(ПРИЧИНА$10/05~!От,Lab(!Деепр,ОГОНЬ$125~!Тв))
ГОРЕТЬ {Глаг} Oper00(!Деепр, Magn_a~ЖЕЛАНИЕ$1503(!Тв))

Table 2. Semantic description of the verb *ИДТИ* [to go]

ИДТИ
Caus(#,IncepLab(МЕСТО$111~!Им,ДЕЙСТВИЕ$13~!Куда))
Caus(#,IncepUsor(ДЕНЬГИ$1114~!Им,!наВин))
Caus(ЖЕНЩИНА$11413119~!Им,Incep~Loc(!Им,ЗАМУЖ$81\НЕКТО$1141~!заВин)))
Caus(КОРИДОР$11321~!Им,Func(ВОЗМОЖНОСТЬ$10/08~ХОД$13304(!Где,!поДат,!Откуда,!Куда)))
Caus(НЕКТО$1141~!Им,IncepCopul(#,СПЕЦ$114132~!вКого))
Caus(НЕКТО$1141~!Им,IncepOper01(!Им,ДЕЙСТВИЕ$13~!Куда))
Caus(Oper01(!Им,ХОД$13304(!Где,!поДат,СЛЕДОМ$3(!заТв),!Откуда,!Куда)),Incep(!Инфин)))
Func(ДЕЙСТВИЕ$13\ОСАДКИ$112/15\РЕЧЬ$132011~!Им)
IncepFunc(СОБЫТИЕ$10/1\ВРЕМЯ$14\СИЛА$13422\ПУТЬ$112431~!Им)
Oper00(#,СОГЛАСИЕ$121127)
Oper00(ОДЕЖДА$11136~!Им,СООТВЕТСТВИЕ$10/031(!Дат))
Oper01(МЕСТО$111~!Им,ХОД$13304(НЕЧТО$1~!Где,!поДат,СЛЕДОМ$3(!заТв!),!Откуда,!Куда))
Oper01(ЧАСЫ$1113231~!Им,РАБОТА$1311)
Caus(Oper01(НЕКТО$1141~!Им,ХОД$13304(!наПред,!поДат,!Откуда,!Куда)),Incep(!Инфин))

valences cease to be oxygen and hydrogen but create the new entity - water. From a computer science point of view attached words are the *arguments* which are used by the attaching word for producing a new construction, whose semantics may differ considerably from one of its components. Any developed language has a functional nature; and only function superposition is adequate to the sentence structure of such a language. The functional nature of the language is especially clearly and simply manifested in the computation of semantic-grammatical types of prepositional expressions, which is well illustrated by the following examples constructed by a semantic analyzer:

@Когда На(@Пред зорьке(@ОНА$5@Пред утренней))
@Когда В(@(Вин погоду($1~@ОНА$5@Вин другую))
@Откуда Из($12~@Род деревни)
@Откуда Из($18~@Род полета)
@Откуда Из($14/0~@Род живота)
@ИзКого Из($141~@ОНИ$5@Род крестьян)
@Откуда Из($113~@Род кастрюли)
@Почему Из($150~@Род уважения)
@Откуда Из($12~@Род берегов)
@Почему Из($150~@Род любви)
@Когда Ко($17~@Дат дню(@Род рождения))
@Откуда@Какой Со($14~@Род слона)
@Откуда С($12~@Род горы)
@Почему С($15~@Род горя)
@Как С($15~@Тв уважением)
@Откуда С($12~@Род запада)

Let's clarify these expressions by the statement *@Почему С($15~@Род горя)* [out of despair]. Preposition "С" for class $15 of nouns (a noun class conventionally called Psychology) in a genitive case produces a semantic-grammatical type *Почему* [*Why*]. It occurs in the following way. The description of the word *"ГОРЯ" [mistery]* will be presented in the form of three morph-semantic alternatives (feeling, run the fire effect and show a strong desire) after the preliminary processing. See Table 1.

One of the preposition "С" alternatives (the overall number of alternatives is 34) looks like follows:

С {Предл $15~@Почему} Caus_y(ПСИХИКА$15~!Род, #)

This alternative can interact with the first alternative of the word "*Горя*". The meaning "*Почему*" calculates as a result of interaction of these alternatives.

The quality of any semantic analyzer may be evaluated depending on how it computes semantic-grammatical meaning of the prepositional-case forms. But even if it computes these meanings with absolute accuracy it doesn't mean that it is able to manage all semantic analysis details. The substitution of the case form as an argument of attaching word sometimes requires very scrupulous

Table 3. Example of translation

Caus(ЧЕЛОВЕК$141~!Им,InсepOper01(ЧЕЛОВЕК$141~!Им,ДЕЙСТВИЕ$18~!Куда)) – !Им (класса ЧЕЛОВЕК) делает так, чтобы начать совершать [ДЕЙСТВИЕ] !Куда

Table 4. Sentence transformation

Бумага БУМАГА {Сущ Неодуш $1127~@ОНА$5@Им} $1127(!Род,!Из) идет *(в описании из-за его громоздкости оставлена лишь одна альтернатива)* ИДТИ {Глаг} Caus(#,IncepLab(МЕСТО$11~!ОН$5\!ОНА$5\!ОНО$5,ДЕЙСТВИЕ$18~!Куда)) в В {Предл @вВин @вПред @вКого @вОНИ$5@Им} $71(!Вин\!Пред\!Кого\!ОНИ$5@Им) В {Предл $18~@Куда} Direkt_y(#,ДЕЙСТВИЕ$18~!Вин) , В {Предл $12~@Куда} Direkt_y(#,ВНУТРИ$30003~ПРИРОДА$12~!Вин) Etc. (There are 33 alternatives). переработку ПЕРЕРАБОТКА {СущНеодуш$18~@ОНА$5@Вин} $1827(!Тв,!Род, !вВин\!наВин) Symbol: .

computing. The precision of semantic analysis directly depends on the quality and completeness of semantic vocabulary.

The proposed approach allows attaining an analysis as accurate as possible by a dictionary enlargement while keeping its structure. And it means that first, the semantic analyzer becomes independent from the dictionary and, second, one can realize a smooth transition from the language semantics to its pragmatics by extending the dictionary.

3 Semantic dictionary

An entry of the computer semantic dictionary contains an entry word and its interpretation in the semantic language. Many words, as a rule frequently used words, have more than one interpretation. Perhaps, the most polysemantic word is the verb *ИДТИ* [to go], whose reduced semantic description has the form presented in Table 2.

Each alternative represents an expression in the semantic language, and can be rather easily translated into Russian (at least into broken Russian). An example can be seen in Table 3.

The main task of the semantic analyzer in investigating a particular sentence is the proper selection of the alternative. This choice is determined by the class and the case forms of the arguments.

4 Preliminary text processing

The word-by-word processing of each text sentence is performed at the stage of the preliminary processing. The first task of this stage is to construct the morph-semantic alternatives. They are independent of each other. These alternatives describe each sentence's word form. The second task is to compute the semantic-grammatical type of each alternative included in the word description. These transformations are necessary for the proper work of the semantic analyzer itself. For example, as a result of these transformations the sentence *"Бумага идет в обработку" [The paper goes to processing]* has been presented in the form presented in Table 4.

After the preliminary processing, the description of each word in the sentence represents a set of alternatives in the identical form, each of which consists of two parts: morphological and semantic. The morphological part (in curly braces) contains information about entities to which this alternative can be attached; the semantic part includes information about components which can be attached to this alternative. All this information is necessary and sufficient for the proper choice of alternatives and their proper combining in the superposition construction.

Table 5. Adjectives *любознательный [inquisitive or investigative]* и *любопытный [curious or inquisitive]*

ЛЮБОЗНАТЕЛЬНЫЙ {Прил$141~@ОН$5@Им} EmCaus_a1(!%1,Hab(!%1,ЗНАНИЕ$151542))
(Такой человек, который склонен приобретать знания)
[Such person who is declined to acquire knowledge, *investigative person*]
ЛЮБОПЫТНЫЙ {Прил @ОН$5@Им} Caus_a1(!%1,Oper02(!Для,ЛЮБОПЫТСТВО$15151(!Тв)))
(Такой, который вызывает любопытство)
[Such person which causes curiosity]
ЛЮБОПЫТНЫЙ {Прил $141~@ОН$5@Им} EmOper02_a1($141~!%1,ЛЮБОПЫТСТВО$15151)
(Такой человек, который склонен проявлять любопытство)
[Such person who is declined to manifest curiosity]
Любознательный случай. [investigative incident]
СЛУЧАЙ {Сущ $10/11~@ОН$5@Им $10/11~@ОН$5@Винн} $10/11(!сТв\!уРод\!Среди\!Где)

The relation between an adjective and a noun is impossible due to the inconsistency of semantic classes $141 and $10/11. The result is the broken text.

Любознательный человек. [curious or investigative person] Here is ambiguity!
ЧЕЛОВЕК {Сущ Одуш $141~@ОН$5@Им} $141(!Род)
The component *$141~@ОН$5@Им* exists between the adjective and the noun. The result is:
EmCaus_o1(ЧЕЛОВЕК$141,Hab(ЧЕЛОВЕК$141,ЗНАНИЕ$151542(!вПред)))

Любопытный случай. [curious incident]
The relation exists only for the first alternative. The result is:
Caus_o1(СЛУЧАЙ$10/11,Oper02(!Для,ЛЮБОПЫТСТВО$15151(!Тв)))

Любопытный человек. [curious person]
The weaker relation (only by case) exists for the first alternative (compared to the previous consideration). The result is:
Caus_o1(ЧЕЛОВЕК$141,Oper02(!Для,ЛЮБОПЫТСТВО$15151(!Тв)))
The stronger relation (both by class number and case) exists for the second alternative. The result is:
EmOper02_o1(ЧЕЛОВЕК$141,ЛЮБОПЫТСТВО$15151)
The sentence *Любопытный для меня человек [There is the curious person for me]* has no ambiguity:
Caus_o1(ЧЕЛОВЕК$141,Oper02(Я,ЛЮБОПЫТСТВО$15151(!Тв)))

Precise semantic meanings of such phrases as *Красная смородина [Red currant]*, *Красный партизан [Red partisan], Красный цветок [Red flower],* etc. are calculated in the same way.

5 Semantic Analysis Essence

At the stage of the semantic analysis, the selection of necessary morph-semantic alternatives and their combination into a single structure are carried out. In our example, the morphological part of the description of the word *ПЕРЕРАБОТКА($18~@ОНА$5@Вин)* [PROCESSING] is used by the semantic analyzer as a means for the choice of that alternative of preposition *В*, whose semantic description part contains the same class and case, i.e. the alternative *{Предл $18~@Куда} Direkt_y(#,ДЕЙСТВИЕ$18~!Вин)* of *В* which produces the semantic-grammatical type *$18~@Куда.* The word *БУМАГА* [*PAPER*] contains the morphological description - *($11~@ОНА$5@Им)*. This description with the semantic type *$18~@Куда* of the phrase "*в переработку*" let the semantic analyzer select the alternative for the verb *ИДТИ* [*TO GO*]:

ИДТИ {Глаг} Caus(#, IncepLab(МЕСТО$11~!ОН$5\!ОНА$5\ОНО$5, ДЕЙСТВИЕ$18~!Куда)).

The substitution of the agreed by class and case arguments of the verb *ИДТИ* [*TO GO*] into its semantic formula, will produce the translation of the source sentence in the semantic language:

Caus(#,IncepLab(БУМАГА$11,ПЕРЕРАБОТКА$18)).
(Кто-то делает так, чтобы бумага начала подвергаться действию переработки)
[Someone ha begun a process regarding the paper]

Let us consider another example illustrating the semantic analyzer work.
Let's take adjectives *любознательный [inquisitive or investigative]* и *любопытный [curious or*

inquisitive]. See Table 5. All necessary explanations are presented in this table.

6 Conclusion

The result of the semantic analysis is the text in the formal semantic language which is the superposition of basic functions and base concepts. When you deal with real tasks it is necessary to use pragmatic analysis of text in connection with particular situations and subject domains, of course. In this case someone has to build some mapping of the text for the used model of task and reality. The text representation in the form of superposition of functions allows making this analysis a direct extension (or further specification) of the semantic analysis: basic concepts are transformed into active objects and basic functions are transformed into operations defining object interaction.

References:

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999, - 513 p. (ISBN: 0-201-39-829-X)

[2] Christopher D. Manning and Hinrich Schuetze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 2000, - 680 p. (ISBN: 0-262-13360-1)

[3] NTCIR Workshop http://research.nii.ac.jp/ntcir/workshop/

[4] TREC http://trec.nist.gov/

[5] Melchuk I.A. *The Experience of The Theory of Linguistic Models "Meaning-Text"*. Moscow, 1979. (in Russian)

[6] Apresyan U.D. *The Lexical Semantic* (selected works). Vol 1. Moscow, 1995. (in Russian)

[7] Chomsky N. Aspects of the Theory of Syntax. Cambridge (Mass.), 1965

[8] Vinograd T.A. *The Program Understands Natural Language*. Moscow, 1976. (in Russian)

[9] Minsky, N., and Rozenshtein, D. A Law-Based Approach to Object-oriented Programming. *SIGPLAN Notices vol.22(12),* October 1987.

[10] Tuzov V.A. *The Computer Semantics of Russian.* Proceedings of Dialog-2001: http://www.dialog-21.ru/ Archive/ 2001/ volume2/ 2_53.htm (in Russian)

[11] ReadRace Tuzov V.A. *Languages for Knowledge Representation*. St.Petersburg, St.Petersburg State University Press, 1990. (in Russian)

[12] Tuzov V.A. *Computer Linguistics*. St.Petersburg, St.Petersburg State University Press, 1998. (in Russian)

[13] Korhov A.V., Korhova O.V. The Algorithm of Resolving the Task of Automatic Search With Using The Method of Formalization of The Russian Language. *Deposits collection in VINITI*, Moscow: № 70-B01. (in Russian)

[14] Korhova O.V. The Method of Formalization of The Russian Language in The Tasks of Building Knowledge Bases and Annotations, *In Proceedings of the XXXII Scientific Conference of the PM-PU Faculty of St.Petersburg State University*. St.Petersburg, 2001. (in Russian)

[15] Korhov A.V. The method of building question-answering system using the mathematical formalization of Russian language, *In Proceedings of the XXXII Scientific Conference of the PM-PU Faculty of St.Petersburg State University*. St.Petersburg, 2001. (in Russian)

[16] Lezin G.V., Boyarski K.K., Kanevski E.A., Popova A.I. The Text Analysis: Representation and Processing of Conceptual Information, In *Proceedings of International Workshop Dialog -97 on Computer Linguistics and its Applications*. Yasnaya Polyana, 1997, pp. 170-174. (in Russian)

[17] Lezin G.V., Mamedniyazova N.S. About Representation of Semantics of Conceptual Models in Knowledge Bases, In *Proceedings of International Workshop Dialog-2000 on Computer Linguistics and its Applications*. Protvino, 2000, Vol 2, pp. 235-242. (in Russian)

[18] Kanevski E.A., Klimenko E.N., Tuzov V.A. About One Approach to Classification of Adjectives, In *Proceedings of International Workshop Dialog-2000 on Computer Linguistics and its Applications*. Protvino, 2000, Vol 2, pp. 162-167. (in Russian)