Verb Taxonomy: from Theoretical Lexical Semantics to Practice of Corpus Tagging

Galina I. Kustova, Olga N. Lashevskaja, Elena V. Paducheva, Ekaterina V. Rakhilina

VINITI and V. V. Vinogradov Institute of Russian Language,
Russian Academy of Sciences, Moscow

One of the main advantages of the Russian National Corpus is its semantic tagging based on an extensive semantic classification of the lexicon – separate for nouns, verbs, adjectives, adverbs and numerals (see Kustova e.a. 2005). Semantic annotations, combined with morphological tagging, make it possible to answer queries that concern not only separate words and word forms but also semantically characterized classes of words, as well as constructions with components presented as morphologically and/or semantically described classes.

Keywords: semantic tagging, verb, taxonomy, ontological classification, word-sense disambiguation, semantic filters, Russian

1. Introduction

Semantic tagging of the Russian National Corpus (RNC, www.ruscorpora.ru) is being carried out with the help of the database, where every dictionary meaning of a word is assigned a set of characteristics along the following parameters:

- taxonomic class: 'persons', 'spaces', 'texts', etc. (for nouns); 'motion', 'location', 'emotion' etc.(for verbs); 'speed', 'duration', 'place' etc. (for adjectives and adverbs);
- mereological class (for nouns): names of 'parts', 'sets' etc.;
- topological class (for nouns): names of 'containers', 'horizontal surfaces', etc.;
- causative/non-causative (for verbs);
- positive and negative evaluation (for all parts of speech);

• derivational class: diminutives (for nouns), prefixal (for verbs), derived from nouns (for adjectives), etc.

At present our database consists of 375000 entries, of which 33000 are entries for verbs.

The classification in the database follows the multi-facet principle: each parameter constitutes a separate principle of division. In other words, we have several classifications (some of them hierarchical) independent of one another. In this paper we deal only with verbs. In Section 1 we present the taxonomic classification of verbs in the database. In Section 2 we describe the application of this classification to semantic tagging of the Corpus, paying attention to ambiguity that arises and to disambiguation devices used to overcome it.

2. Taxonomic classification of verbs

2.1. General

Modern semantics distinguishes two main taxonomic classifications of verbs: one may be called **ontological** (or **thematic**), the other is **actional** (it is also called **aspectological**).

Among ontological classifications of verbs, the most developed one is that suggested by Beth Levin (1993) for English verbs. This classification, though having indisputable merits, was carried out in the ideology of transformational grammar: every class is substantiated by its participation in transformations of different kinds, such as diathesis alternations, locative shift, morphological derivations and so on. This feature of Levin's classification is discussed in Baker & Rupperhofer (2000), where it is compared with the classification in FrameNet. Their conclusion is that concentration on formal procedures sometimes leads to classes irrelevant from purely semantic point of view.

In our ontological classification, we appeal directly to the meaning of a word. It goes without saying that we wanted to get classes that would be linguistically relevant, i.e. classes of words that would be similar in their linguistic behavior. But we relied upon semantics, starting from the assumption that semantically identifiable classes might be a reliable basis for grammatical predictions. Also, other things being equal, semantic classes tend to be more universal.

The actional classification is a ramification of the famous Vendler's classification according to which verbs are divided into four classes – actions, activities, processes and states. Actional class determines different details of linguistic behavior of a verb (aspect, voice, co-occurrence with modifiers of time and purpose, combinability with aspectual verbs etc.) and is one of the central notions in modern theoretical grammar studies.

Thematic and actional classifications are independent of one another (Paducheva 2004b). On the one hand, there are mental, perceptive, emotional, volitional, physical, physiological, social actions and states; on the other hand, mental verbs may belong to the class of states (*znat'* 'know', *pomnit'* 'remember'), actions (*rešat'* – *rešit'* 'to solve') and activities (*razmyšljat'* 'reflect').

In this paper we discuss only thematic, i.e. ontological classification. Actional classes won't be in the focus of our attention. Thus, verb taxonomy is understood, in this paper, as a thematic classification of verbs.

Let us begin with some general remarks about thematic classes.

Thematic classification, according to its very nature, is such that a word, even a non-ambiguous word, may belong to several different classes. Class attribution of a word (more precisely, a word taken in one of its meanings, a **lexeme**), is determined by the corresponding **semantic component** in its meaning. For instance, *videt'* 'see', *slyšat'* 'hear', *podgljadyvat'* 'watch furtively' have a common semantic component 'perception', and, thus, belong to the class of PERCEPTION; *bit'* 'beat', *rezat'* 'cut' have a common semantic component 'impact' and belong to the class of IMPACT. (Presumably, these components predict some feature(s) of behavior of the lexeme). But the meaning of a lexeme may include several equally conspicuous semantic components, and thus it may be identified as belonging to several different classes. For example, *ubedit'* 'convince' is a verb of SPEECH and, at the same time, of influencing volitional and mental states;

ogljanut'sja 'look back' is a verb of CHANGE OF POSITION and PERCEPTION; *zastat'*, *zastič* 'take unawares' – MOVEMENT and PERCEPTION. The verb *skryt'sja* 'escape' has the same two components, though in a different configuration. The verb *oblokotit'sja* 'lean one's elbows on a surface' belongs to the class of SPATIAL CONFIGURATION and CONTACT & SUPPORT. A widespread conjunction of semantic components (and, therefore, thematic classes) characterizes the verb *napolnit'* 'load': PUTTING and CHANGE OF STATE OR PROPERTY; the verb *zabit'* <*gvozd'*> 'drive <a nail>' – PUTTING and IMPACT. The verb *plakat'* 'weep, cry' belongs to two classes – SOUND and PHYSIOLOGY.

The fact that a non-ambiguous word (or a lexeme) belongs to two different thematic classes characterizes its meaning and is not a drawback of the classification. In fact, *napolnit'* 'load', which is assigned to the classes PUTTING and CHANGE OF STATE OR PROPERTY, has a more sophisticated semantics than *položit'* 'lay', which expresses only 'putting', and *namoknut'* 'get wet', which expresses only 'change of state'.

On the other hand, a verb can be **ambiguous**, and then its multiple class attribution reflects its multiple meanings.

Example 1: *zametit'* =

(i) 'notice', verb of PERCEPTION;
(ii) 'make a remark', verb of SPEECH.

Example 2: *videt'* =

(i) 'see', verb of PERCEPTION;
(ii) 'believe', verb of MENTAL SPHERE (*On vidit vo mne sopernika* 'He considers me to be a rival').

Ambiguity is a characteristic feature of the lexicon of any language. At the same time, used in context an ambiguous word normally (i.e. if we exclude pun) has only one meaning. Therefore, the semantic annotation of corpus texts presupposes disambiguation. This task is discussed in section 2.

What information is deduced from the verb's ontological class?

1. First of all, thematic class influences the argument structure of the verb – the set of its participants (theta roles). The fact is that verbs of the same thematic class usually refer to the same typical situation. Verbs of creation must have the participant Result, verbs of (bounded) movement must have Initial point (cf. *ujti* 'leave') and Final point (cf. *prijti* 'arrive'). Verbs of speech imply Addressee and Text; verbs of selling and buying presuppose Agent, Counteragent, Goods, Money; verbs of sound imply Sound Emitter, Observer (off stage) and Sound (as an incorporated participant, see Jackendoff 1990: 61; Paducheva 2004a: 57-58).

Thematic classes "are at the heart of the area of linguistics called argument structure: the study of the possible syntactic expressions of the arguments of a verb" (Levin & Rappaport Hovav 2005:1). The issue is illustrated with the classic example from Fillmore 1977:

(1) a. The boy *broke* the window with a ball;
b. The boy *hit* the window with a ball.

(2) a. The window *broke*;
b. *The window *hit*.

Though (1a) and (1b) may describe the same situation, (1a) has a lexico-syntactic correlate (2a), while (1b) does not. And, what is important, this is not an isolated pair of verbs: *hit* and *break* are representatives of their thematic classes: *break* belongs to verbs of PHYSICAL IMPACT and *hit* is a verb of CONTACT. Contact does not necessarily imply effect (and change of state): it is possible that the ball hit the window but the window did not break.

Analogous differences are found in many languages, so it is reasonable to assume that different languages have similar thematic classes.

2. Thematic class predicts, to a reasonable degree, the **derivational potential** of a word – a set of its possible derived meanings. For example, mental derivatives of perception verbs (*Videl v nej tol'ko sopernika* 'He saw in her only a rival'); existential meaning of verbs of non-directed movement (*V prudu plavali utki* 'There were ducks swimming in the pond'); speech meaning of verbs of emotion (*Gosti vosxiščalis' vidom iz okna* 'The guests admired the view out the window'). A paradigm of derived meanings of sound emission verbs is presented in Atkins et. al.(1988). A semantic derivation paradigm for verbs of *spray-load* class is given in Paducheva & Rozina (1993).

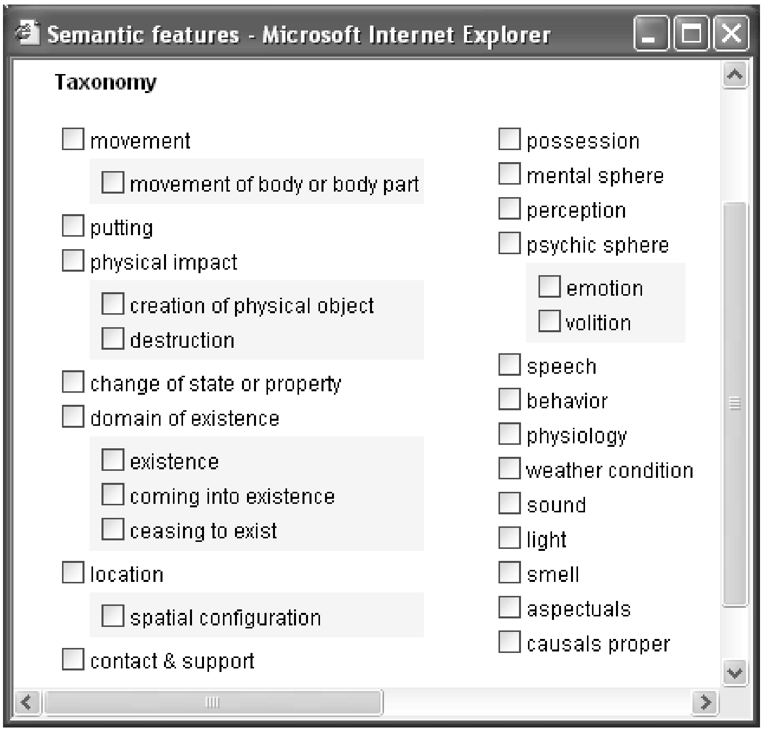
3. Thematic class may predict **referential status** of a participant. For example, the object of a CREATION verb is, normally, an indefinite NP: *postroit’ <dom>* ‘build <a house>’, *napisat’ <poemy>* ‘write a poem’. (Hence the genitive object of creation verbs in negative sentences in Russian.)

Before we go down to the list of thematic classes, the notion of a **constructional component** must be introduced (Paducheva 2004a: 45). Constructional components should be delineated in order to substantiate thematic classifications: two words that are differentiated solely by a constructional component may belong to one and the same thematic class. Cf. such constructional components as **causation** (*videt’* ‘see’ – *pokazat’* ‘show’; *prijti* ‘come’ – *prislat’* ‘send’ belong to the same thematic class), **modality** (*vpustit’*‘allow to enter’ belongs to the same class as *vojti* ‘enter’), **negation** (*otkazat’sja* ‘refuse’ belongs to the same class as *soglasit’sja* ‘agree’), **inception** (cf. *zacvesti* ‘start to flower’ and *cvesti* ‘flower’).

2.2. Ontological (= thematic) classes of verbs

Thematic classes of verbs now searchable in the RNC are shown in Fig.1. Below we discuss thematic classes of verbs one by one, providing each class with its linguistic substantiation.

1. MOVEMENT (*bežat’* ‘run’, *djorgat’sja* ‘twitch’, *brosit’* ‘throw’, *nesti* ‘carry’). This is one of the largest classes (more than 2000 words). There are different subclasses inside this class. Some subclasses can be identified (for the sake of information retrieval) by means of prefixes (*vy-, iz-, s-, ot-* for elative verbs; prefixes *v-, do-, pri-, za-* for lative verbs). Some subclasses of movement verbs are traditionally used in the Russian grammar because they are of direct relevance to morphology (such as verbs of non-directed movement, *xodit’*‘walk’, *begat’*‘run’, *ezdit’*‘ride’, *letat’*‘fly’ etc.).



Semantic features - Microsoft Internet Explorer

Taxonomy

- □ movement
  - □ movement of body or body part
- □ putting
- □ physical impact
  - □ creation of physical object
  - □ destruction
- □ change of state or property
- □ domain of existence
  - □ existence
  - □ coming into existence
  - □ ceasing to exist
- □ location
  - □ spatial configuration
- □ contact & support
- □ possession
- □ mental sphere
- □ perception
- □ psychic sphere
  - □ emotion
  - □ volition
- □ speech
- □ behavior
- □ physiology
- □ weather condition
- □ sound
- □ light
- □ smell
- □ aspectuals
- □ causals proper

Fig. 1. Thematic classes of verbs

1.1. MOVEMENT OF BODY OR BODY PART (*leč* 'lie down', *vstat'* 'stand up', *nagnut'sja* 'bend down'; *sognut'* < *ruku v lokte*> 'bend < one's arm at the elbow>'). This is the only subclass inside MOVEMENT verbs that is distinguished as a separate class in the RNC. Moving your own body part is something different from moving external physical objects: *sognut' ruku* <*v lokte*> 'bend one's arm' is not the same as *sognut' kočergu* 'to bend a poker'. Movement of body parts presupposes inner causation (according to Wierzbicka 1980), or psychological causation according to Podlesskaya & Rakhilina (1999).

Class 1.1 is restricted as to taxonomic class of its object: not only must it be a body part – it must be a body part of the Agent – *položit'* in *položit' kurinye nožki na skovorodku* 'to put the chicken legs into a frying-pan' is not a verb of MOVEMENT OF BODY OR BODY PART.

It is important to compare class 1.1 with class 6.1 SPATIAL CONFIGURATION (containing verbs *stojat'* 'stand', *sidet'* 'sit', *ležat'* 'lie', *viset'* 'hang', their synonyms: *vossedat'* 'sit (as) on the throne', and derivates: *posidet'* 'sit for a bit'). Class 6.1 contains only non-causative verbs, while class 1.1 includes both non-causatives and causatives (*nagnut' golovu* 'bend one's head’). This means that in class 1.1 causation is treated as a constructional component. Note that in Wierzbicka 1980 intransitive verbs of class 1.1, such as *vstat’* ‘stand up’, are treated as causatives; in fact, the meaning of *vstat’* implies an incorporated participant Body: *vstat’* means ‘cause your body to be in a vertical position’. Class 1.1 differs from 6.1 also in that verbs in 6.1 constitute aspectual pairs (*leč* – *ložit’sja* ‘lie down’), which is normal for verbs of movement, while verbs in 6.1 are imperfectiva tantum.

2. PUTTING (or PLACEMENT OF OBJECT according to Levin 1993). This class consists mostly of transitive verbs (*položit’* ‘put’, *vložit’* ‘put in’, *sprjatat’* ‘hide’). Not all transitive verbs of movement belong to the PUTTING class. The semantics of verbs *nesti* ‘carry’, *vezti* ‘drive’, for instance, don’t correspond to that of PUTTING class verbs: *nesti* and *vezti* imply movement of the subject, not only of the object. Another example: the verb *brosit’* ‘throw’ doesn’t belong to PUTTING class because ‘to throw’ doesn’t necessarily imply ‘to put’.

Verbs of PUTTING are semantically close to verbs of PHYSICAL IMPACT, class 3. For example, *zasunut’* ‘slip into’ is a verb of putting; it has two participants different from the Agent: “What?” and “Where?”, while the verb *zabit’* (in the context *zabit’ gvozd’ v stenu* ‘drive a nail into the wall’) belongs to two classes: PUTTING and PHYSICAL IMPACT. The verb *zaplombirovat’* ‘stop a tooth’, which has the same two participants (one of them incorporated, namely, stopping) is included in the class PHYSICAL IMPACT, the tooth being treated, in the first place, as the object of impact, not as the place for a stopping (cf. the verb *zaminirovat’ <gavan’>* ‘mine <the harbour>’).

3. PHYSICAL IMPACT (*bit’* ‘beat’, *kolot’* ‘prick’, *vytirat’* ‘wipe’). According to several linguistic sources, this class contains many different subclasses. Thematic classification of the RNC brings forward two subclasses inside the class of physical impact:

3.1 – CREATION (*vykovat’* ‘forge’, *sšit’* ‘sew’);

3.2 – DESTRUCTION (*vzorvat’* ‘explode’, *sžeč* ‘burn down’, *zarezat’* ‘slaughter’, *ubit’* ‘kill’).

There is an important subclass of CREATION verbs – verbs OF IMAGE CREATION, such as *narisovat'* 'draw', *vyšit'* 'embroider', *nadpisat'* 'inscribe' (see Levin 1993). These verbs are mentioned in Fillmore (1977), because of the double semantic role of their syntactic object. In (3) *Maša* denotes either Maša herself or her portrait:

(3) I painted Maša.

Class 4 CHANGE OF STATE OR PROPERTY verbs will be discussed later – for the reasons that will become clear in a while.

5. EXISTENCE. Three subclasses are distinguished in class 5:

5.1. EXISTENCE (*žit'* 'live', *proisxodit'* 'happen', *vodit'sja* 'be found', *suščestvovat'* 'exist')

5.2. COMING INTO EXISTENCE (*vozniknut'* 'arise', *rodit'sja* 'be born', *sformirovat'* 'form', *sozdat'* 'create')

5.3. CEASING TO EXIST (*likvidirovat'* 'liquidate', *iskorenit'* 'root out', *vymeret'* 'die out', *issjaknut'* 'run dry' <about moisture>, *izgladit'sja* 'to be blotted out' <of one's memory>).

The class of COMING INTO EXISTENCE verbs includes such verbs as *zaroždat'* 'generate', *roždat'* 'give rise to', *plodit'* 'produce', *razžigat'* 'inflame', *sozdavat'* 'set up', i.e. verbs of causing existence of a situation, not a physical object.

There is a close affinity between CREATION verbs, class 3.1, and the subclass 5.2 of verbs of EXISTENCE. A problem arises in connection with the borderline between verbs of class 5.2 with the meaning 'cause to come into existence' (*sformirovat'* 'form', *sozdat'* 'create'), and class 3.1. In its prototypical manifestations creation is not reducible to 'begin' and 'exist'. In fact, semantics of a verb of creation presupposes an activity of a specific kind which leads to the object's coming into existence, (see Wierzbicka's (1980) and Fodor's (1970) arguments about why *to kill* does not mean 'cause to die'). In other words, CREATION verbs are mostly actions. Which is not the case with 'cause to begin to exist' verbs. Here again aspectological considerations interfere with purely semantic divisions. Such verbs as *žit'* 'live' and *rodit'sja* 'be born' belong not only to verbs of EXISTENTCE, but to the class PHYSIOLOGY as well. Still, in their derived use (as in *rodilas' ideja* 'an idea was born') they denote pure COMING INTO EXISTENCE. Note that in the class of COMING INTO EXISTENCE verbs the component INCEPTION cannot be treated as constructional.

6. LOCATION is a class of primary importance ontologically (see, for example Levin & Rappaport Hovav (2005: 79) on centrality of localization component in the theory of semantic roles). In fact, for a material object it is natural to be localized somewhere.This is an axiom that occupies an important place in the sphere of entailment rules of lexical semantics. Still, lexically LOCATION is not a clear cut class; it is represented by a stylistically deficient *naxodit'sja* 'be situated', highly ambiguous *byt'* 'be' and by secondary uses of verbs of other classes, e.g. by SPATIAL CONFIGURATION verbs in their bleached meanings, as in *Dom stoit na gore* 'The house stands on a hill'.

The LOCATION class is sufficiently numerous – some two dozens of verbs; among them verbs with an incorporated participant Time interval: *nočevat'* 'pass the night', *zimovat'* 'pass the winter'.

In the LOCATION class 'causation' is not a constructional component: the causative of a location verb doesn't belong to the same class as the verb itself; so *položit'* 'put', a causative of *ležat'* 'lie', is not a verb of location, it is a verb of PUTTING, class 2.

6.1. SPATIAL CONFIGURATION; verbs of this class are also called STANCE verbs; in Rakhilina (2000) they are called verbs of POSITION: *stojat'* 'stand', *ležat'* 'lie', *sidet'* 'sit', *viset'* 'hang'. Note that *prislonit'sja* 'lean against', *oblokotit'sja* 'lean one's elbows on a surface' also belong here. Verbs of this class have a common set of semantic valencies (on valencies of SPATIAL CONFIGURATION verbs see Apresjan 2006): Agent (*Ivan sidit* 'Ivan is sitting'), Location (*stojat' na uglu* 'stand on the corner')', Supporting Body Part (*stojat' na odnoj noge* 'stand on one leg'), and Direction (*sidet' licom k stene* 'sit with one's face to the wall').

A verb of SPATIAL CONFIGURATION class may have the meaning of LOCATION; in which case the participant LOCATION becomes obligatory:

(4) a. *Počemu ty stoiš? Sjad'* 'Why are you standing? Sit down' [non-obligatory participant Location];

b. *Cerkov' stoit na gore* 'The church stands on the hill' [obligatory participant Location].

7. CONTACT & SUPPORT (*kasat'sja* 'touch', *obnimat'* 'embrace', *oblokotit'sja* 'lean one's elbows on a surface', *opirat'sja* 'lean against smth').

This is a big class, more than one hundred words. It is linguistically important, because here a thematic component implies stativity, an actional feature (Vendler's STATE). The borders of this class are not quite clear: contact is a prerequisite and at the same time a consequence of many different and heterogeneous situations. In the corpus only the verbs for which "contact" is an assertive component are included in this class.

Classes 8. POSSESSION, 9. MENTAL SPHERE, 10. PERCEPTION, 11. 1.EMOTION, 11.2. VOLITION, 12. SPEECH, 14. PHYSIOLOGY (*est'*, *pit'*, among others) are commonly accepted and need no comments.

In thematically clear-cut classes causativity and inceptivity function as true constructional components; for example, *pokazat'* 'show', a causative verb, belongs to the same class (PERCEPTION) as *uvidet'* 'see'; *zabolet'* 'fall ill', an inceptive verb, belongs to the same class (PHYSIOLOGY) as *bolet'* 'be ill'.

Class 13. BEHAVIOR contains such verbs as *šalit'* 'be naughty', *priveredničat'* 'be fastidious', *potvorstvovat'* 'connive'. This is a new class; it is only recently that it was deemed linguistically relevant. Some verbs of this class are conceived as "verbs of interpretation" in Apresjan (2004b).

Class 15. WEATHER, contains verbs with natural forces as a subject (as in *buševala burja* 'the storm raged', *dul veter* 'the wind blew').

Classes 16. SOUND, 17. LIGHT, 18. SMELL are represented in Levin (1993) as subclasses of EMISSION verbs (also including such verbs as *dymit'sja* 'smoke', *krovotočit'* 'bleed', *puzyrit'sja* 'bubble', *penit'sja* 'foam').

The last two small classes of the thematic classification – 19. ASPECTUALS (e.g. *nachat'* 'start', *prodolzhat'* 'continue', *prekratit'* 'stop') and 20. CAUSALS PROPER (*vyzvat'* 'cause', *privesti k* 'bring to') are unproblematic.

Let us now return to class 4, CHANGE OF STATE OR PROPERTY (*povzroslet'* 'become adult', *razbogatet'* 'become rich', *rasširit'* 'widen', *ispachkat'* 'dirty up'). Emotional, mental and physiological states (e.g. *zabolet'* 'fall ill') are provided with thematic classes of their own, so that CHANGE OF STATE OR PROPERTY is a kind of default class: verbs formed from adjectives belong to this class; but verbs that denote a change of a certain specific state belong to more specific classes, such as MOVEMENT, EMOTION, POSSESSION, MENTAL SPHERE, CONFIGURATION and are not included here. There are many different states and properties denoted by adjectives that may change: color (*poželtet'* 'turn yellow'), form (*vytjanut'sja* 'lengthen'), weight (*oblegčit'* 'lighten'), temperature (*oxladit'sja* 'chill'), attribute of a person (*poglupet'* 'grow stupid'), see the taxonomy of adjectives in the RNC.

The class of CHANGE OF STATE OR PROPERTY VERBS is not to be confused with the class of **change of state** verbs, see, e.g., Levin & Rappaport Hovav (2005: 89). The latter term is widely used by linguists of different theoretic orientations. Change of state verbs do not constitute a thematic class in the sense in which verbs of movement or verbs of perception do; the class of CHANGE OF STATE OR PROPERTY VERBS includes only a small part of change of state verbs. Change of state verbs seem to occupy an intermediate position, being neither purely thematic nor purely actional.

The causatives of a CHANGE OF STATE OR PROPERTY verb belong to the same class as the initial verb; for example, *oxladit'* 'chill something.' belongs to the same class as *oxladit'sja* 'chill'.

There is a subclass of CHANGE OF STATE OR PROPERTY verbs derived, formally or semantically, from the comparative form of an adjective and denoting a change of a parameter, such as size, length, velocity, price, etc. (on their aspectual relevance see Glovinskaja 1982: 86). These verbs are called GRADUALS, examples: *uveličit'sja* 'increase', *sokratit'sja* 'decrease', *uskorit'* 'quicken', *podorožat'* 'rise in prise'.

Some verbs of CHANGE OF STATE OR PROPERTY class have a derived meaning of manifestation (of a property); for example, *belet′* = 'to be seen as white'. These are verbs implying the presence of the observer.

2.3. Evaluation

Though our major topic is ontology, a few words should be said about another facet of the verb classification in the RNC database, namely, **positive and negative evaluation**. Semantics of verbs of BEHAVIOR include, more often than not, the component "negative evaluation", as in *vytvorjat′* 'be up to no good', *natvorit′* 'make balls of something', *podsidet′* 'intrigue against', *pilit′* 'pester, lit. saw'. There are many verbs with negative evaluation in the class of verbs of speech (*komkat′* 'crumple up <one's speech>', *mjamlit′* 'mumble', *nyt′* 'whimper', xamit' 'be rude'), verbs of possessive sphere (*obsčitat′* 'cheat', *pičkat′* 'stuff with', *navjazat′* 'thrust <one's opinion>'), and even verbs of perception (*ustavit′sja* 'stare stupidly'). The verb *prozevat′* 'miss, lit. yawn' (as in *prozevat′ rasprodažu* 'miss the sale') differs from *propustit′* 'miss' only in its negative evaluation component.

Many verbs acquire the negative evaluation component when used in their derived meaning: *mazat′* 'lit. oil' in the meaning 'paint badly'; *komkat′* 'lit. crumple a piece of paper' in the meaning 'crumple one's speech', *zavodit′sja* 'get wound up' in the meaning of a psychological state.

The class of interpretation verbs (see Apresjan 2004b; Kustova 2004: 232-241) is not present among RNC ontological classes. It could have been localized in the facet 'evaluation'.

In fact, verbs of interpretation may belong to different thematic classes: *narušit′* 'violate', *oblegčit′* 'alleviate', *pomoč* 'help', *pomešat′* 'hinder', *spasti* 'save' interpret physical actions; *opozorit′* 'dishonour' interprets a psychological state; *preuveličivat′* 'exaggerate' refers to speech. The verb *ošibat′sja* 'be wrong' is thematically (and actionally) ambiguous (see Mel'čuk 2004: 45): in the imperfective it means 'to have a wrong opinion', while in the perfective it can be used in the meaning 'to perform a wrong action'.

The same can be said about the class of manifestation verbs – both such as *belet'* 'be seen white', *zvučat'* 'sound', *paxnut'* 'smell', *gorčit'* 'taste bitter', *žečsja* 'sting' (all characterized by the presence of observer-experiencer), and *plakat'* 'cry', *okamenet'* 'turn into stone' (which are emotion manifestations, Iordanskaja 1972, Apresjan 2004a).

3. Semantic annotation and word sense disambiguation

A semantically tagged corpus makes it possible to verify various linguistic hypotheses concerning semantic and syntactic compatibility of words and forms in texts. Every query addressed to a corpus is a potential **construction** in the sense of Fillmore (Fillmore & Kay 2005, Fried & Östman 2004), i.e. a hypothesis of admissible configuration of words with certain characteristics. Below we deal with configurations of words and word classes defined by a combination of morphological and semantic tags. We intend to demonstrate how the notion of construction contributes to the semantic disambiguation of a verb in a given context. Special attention is paid to the role of taxonomy (of verbs and nouns).

Up till now much research was devoted to the role of the **syntactic context** in verb disambiguation. For example, Levin & Rappaport Hovav (2005) relies upon **taxonomy** of the words constituting the context and its interplay with the taxonomy of a verb itself. In other words,

1) we pay attention to ambiguity expressible in terms of ontological classes of verbs;

2) we consider the taxonomic class (of nouns, adverbs etc.) to be one of the main disambiguating factors – a factor that up till now has remained unnoticed.

Every verb class has semantic prerequisites concerning the thematic class of each argument. A change of the thematic class of the argument may change the thematic class of the verb. This is demonstrated below with verbs which in their primary meaning have a participant necessarily belonging to the thematic class HUMAN.

Example 1 concerns verbs of BEHAVIOR. The subject of such verbs necessarily belongs to the thematic class HUMAN. Some verbs of behavior are non-ambiguous (such as *besčinstvovat'* 'commit outranges', *bujanit'* 'kick up a row', *važničat'* 'put on airs', Paducheva 1996: 149ff); but there are verbs which may mean not only behavior. For these verbs a filter is provided that cancels the tag BEHAVIOR in the context of the non-HUMAN subject. See (5a, 6a) where the verbs belong to the class BEHAVIOR and (5b, 6b) where this is not the case:

(5) a. Petja lomaetsja 'Petja behaves as a poseur';
b. Pribory lomalis' 'The instruments were being broken down'.

(6) a. Petja sovershenno raspustilsja 'Petja got out of hand',
b. Akacija raspustilas' 'The acacia has flowered'.

Example 2 concerns MENTAL verbs, whose subjects belong to the class HUMAN. If the subject doesn't belong to this class the meaning of the verb changes. This can be shown on the example of the verb *znat'* 'know'. *Znat'* 'know' in the negative form, when used in the context of non-HUMAN subject changes its meaning and has a very general meaning of CONTACT:

(7) Ee iznežennye pal'cy ne *znali* igl (A.Pushkin) 'Her delicate fingers didn't know needles'
= 'never were in contact with needles'.

In (8) the thematic class of *znat'* 'know' can be identified as EXISTENCE; (8a) = 'there are no titles and ranks in art'; (8b) = 'successes like this don't exist in the world practice':

(8) a. Iskusstvo ne *znaet* titulov i rangov (RNC) 'The art doesn't know titles or ranks';
b. Mirovaja praktika ne *znaet* podobnyx uspexov (RNC) 'lit. World practice doesn't know such success'.

Example 3: Verbs of speech change their meaning in the context of a non-HUMAN subject. Such shifts are possible for great many verbs (Paducheva 2004: 371): *dokazat'* 'prove', *podtverdit'* 'confirm', *podčerkivat'* 'underline', *izvinjat'* 'excuse', *obeščat'* 'promise', *trebovat'* 'demand' etc.:

(9) a. Kompozicija *podčerkivaet* glavnuju mysl' proizvedenija (RNC) 'The composition stresses the main idea of the piece';
b. Večer *obeščaet* byt' interesnym 'lit. The party promises to be interesting'.

Transferring tags from the semantic database to the the corpus is done in the following way: an ambiguous word in the database has several semantic tagsets which are differentiated by thematic class markers.. The **Primary program** assigns tags to words in the texts automatically; it assigns every occurrence of the word in question all the tags that this verb has in the database. Roughly speaking, this program doesn't know the meaning of a word in a particular context. The second half of the program is called **Semantic filters**. The function of a filter is to isolate the tags of an ambiguous word which correspond to this context and erase all the unwanted tags.

A filter is a combination of lexical, morphological and semantic characteristics determining the context which substantiates (or gives rise to) a particular meaning of the word. All the other tags of a word in the set of examples responding to the filter are then erased. In this way disambiguation is carried out up to the thematic class. The user has the opportunity to formulate queries and receive answers, say, not about the verb idti 'go', but about *idti* in the meaning 'motion' (as in *poezd idet* 'the train goes'); or in the meaning 'take place' (as in *idjot urok* 'the lesson takes place'), etc.

Verb ambiguity is, more often than not, **regular ambiguity**, i.e. polysemy. For example, many verbs of physical impact have a derived meaning in the thematic class SPEECH: *pilit' brevno* 'to saw the log' vs. *pilit' muža* 'to nag one's husband', *rezat' xleb* 'to cat the bread' vs. *rezat' pravdu-matku* 'to cut the truth', *molot' muku* 'to grind the flour' vs. *molot' čuš'* 'to talk nonsense'. Every occurrence of such a verb in the course of automatic tagging, receives two tags, 'impact' and 'speech'. Filters are responsible for the disambiguation.

Several examples of filters are given below. The left part of a filter consists of an ambiguous verb (its meanings, identified by their thematic class, are given in brackets) and a noun with morphological and semantic characteristics. Nouns, according to the semantic classification of the RNC, are subdivided into three major classes: CONCRETE nouns, i.e. names of physical objects, ABSTRACT nouns and PROPER nouns. On the next level of hierarchy concrete nouns are divided into classes HUMAN, ANIMAL, TRANSPORT, STUFF, SPACE, TEXT, FOOD, etc.; abstract nouns are divided into classes MOVEMENT, PHYSICAL IMPACT, …, SPEECH (basically, all verb classes are relevant also for deverbal nouns), as well as TIME, PARAMETER, COLOUR, etc. Here are three filters disambiguating verbs *pilit′*, *molot′* and *s″est′*:

F1. a. ***pilit′*** (impact, speech) + Noun: Acc: CONCRETE (*pilit′ brevno* ‘to saw the log’) → *pilit′* (impact)
b. ***pilit′*** (impact, speech) + Noun: Acc: HUMAN (*pilit′ muža* ‘to nag one’s husband’) → *pilit′* (speech)

F2. a. ***molot′*** (impact, speech) + Noun: Acc: STUFF (*molot′ muku* ‘to grind the flour’) → *molot′* (impact)
b. ***molot′*** (impact, speech) + Noun: Acc: SPEECH (*molot′ čuš′* ‘to talk nonsense’) → *molot′* (speech)

F3. a. ***s″est′*** (physiology; ceasing to exist) + Noun: Acc: FOOD (*s″est′ jabloko* ‘to eat an apple’) → *s″est′* (physiology)
b. ***s″est′*** (physiology; ceasing to exist) + Noun: Nom: STUFF (*ržavchina s″ela mašinu* ‘rust ate the car’) → *s″est′* (ceasing to exist)

In the last example it is not only the thematic class of the noun in the Accusative but also the Dative of Addressee that identifies the occurrence of the verb as belonging to the class SPEECH.

It is not always possible to assign the derived meaning to any of the existing thematic classes. In some cases the tags “metaphor” or “metonymy” are used in order to differentiate meanings, cf. the use of the verb *idti* ‘go’ in *vremja idet* ‘time goes’ or *urok idet*, lit. ‘the lesson goes’.

Thus, the hierarchy of thematic classes may fulfill the role of a co-ordinate system in the intricate network of a word’s meanings: the thematic class serves as a reasonable objective of disambiguation.

In the RNC verb classification semantics was in the foreground. Linguistic relevance of the classes, we hope, should emerge as a consequence of the semantic contiguity between words of the same class – due to the hypothesis of close relationship between combinability of linguistic entities (on the different levels of structure) - and their semantics. The use of the Russian

National Corpus in ongoing linguistic research seems to confirm this hypothesis.

Bibliography

Apresjan Ju. D. (2004a). 'Sinonimičeskij rjad *zameret'* 1.1, *zastyt'* 2.1, *ostolbenet'*, *ocepenet'* 1, *okamenet'* 2'. In: Apresjan V.Ju. e. a. *Novyj ob"jasnitel'nyj slovar' sinonimov russkogo jazyka.* Izd. 2. Moskva-Vena: Jazyki slavjanskoj kul'tury: 348-354.

Apresjan Ju. D. (2004b). 'Interpretacionnye glagoly: semantičeskaja struktura i svojstva'. *Russkij jazyk v naučnom osveščenii*, 1 (7): 5-22.

Apresjan Ju. D. (2006). 'Fundamental'naja klasifikacija predikatov'. In: Apresjan Ju. D. (otv. red.), *Jazykovaja kartina mira i sistemnaja leksikografija*. Moskva: Jazyki slavjanskoj kul'tury. 75-109.

Atkins, B. T., J. Kegl & B. Levin (1988). 'Anatomy of a verb entry: from linguistic theory to lexicographic practice'. *International Journal of Lexicography*, 1 (2): 84–126.

Baker, C. F. & J. Ruppenhofer (2002). 'FrameNet's frames vs. Levin's verb classes'. In: J. Larson & M. Paster (Eds.), *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*: 27-38.

Fillmore Ch. J. & P. Kay (2005). *Construction grammar.* Stanford, CA: CSLI.

Fillmore Ch. J. (1977). 'The case for case reopened'. In: P. Cole & J. M. Sadock (Eds.), *Syntax and Semantics 8: Grammatical Relations*. N. Y. etc.: Acad. Press: 59–81.

Fodor J. A. (1970). 'Three reasons for not deriving 'kill' from 'cause to die''. *Linguistic Inquiry* 1: 429-438.

Fried M. & J.-O.Östman (2004). *Construction Grammar in a Cross-Language Perspective*. Amsterdam: John Benjamins.

Glovinskaja M. Ja. (1982). *Semantičeskie tipy vidovyx protivoposavlenij russkogo glagola*. Moskva: Nauka.

Iordanskaja L. N. (1972). 'Leksikografičeskoe opisanie russkix vyrazhenij, oboznačajuščix fizičeskie simptomy čuvstv'. *Mašinnyj perevod i prikladnaja lingvistika*, vyp. 16. Moskva.

Jackendoff R. S. (1990). *Semantic Structures*. Cambridge etc.: MIT Press.

Kustova G. I., O. N. Lashevskaja, E. V. Paducheva & E. V. Rakhilina (2005). 'Semantičeskaja razmetka leksiki v nacional'nom korpuse russkogo jazyka: principy, problemy, perspektivy'. In: *Nacional'nyj korpus russkogo jazyka: 2003-2005.* Moskva: Indrik. 155-174.

Kustova G. I. (2004). *Tipy proizvodnyx značenij i mexanizmy jazykovogo rasširenija*. Moskva: Jazyki slavjanskoj kul'tury.

Levin B. & M. Rappaport Hovav (2005). *Argument Realization*. Cambridge UP.

Levin B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago UP.

Mel'čuk I. (2004). 'Actants in semantics and syntax I: actants in syntax'. *Linguistics* 42–1: 1-66.

Paducheva E. V & R. I. Rozina (1993). 'Semantičeskij klass glagolov polnogo oxvata: tolkovanie i leksiko-sintaksičeskie svojstva'. *Voprosy jazykoznanija*, 6: 5–16.

Paducheva E. V. (1996). *Semaničeskie issledovanija*. Moskva: Jazyki slavjanskoj kul'tury.

Paducheva E. V. (2004b). 'O parametrax leksičeskogo značenija glagola: ontologičeskaja kategorija i tematičeskij klass'. *Russkij jazyk segodnja*, t.3, Problemy russkoj leksikografii, red. L. P. Krysin. Moskva. 213-238.

Podlesskaya V. I. & E. V. Rakhilina (1999). 'External possession, reflexivization and body parts in Russian'. In: D. L. Payne & I. Barshi (Eds.), *External Possession*. Amsterdam etc.: John Benjamins. 505–521.

Rakhilina E. V. (2000). *Kognitivnyj analiz predmetnyx imen: semantika i sočetaemost'*. Moskva: Russkie slovari.

Wierzbicka A. (1980). *Lingua Mentalis*. Sydney etc.: Acad. Press.