Semantic Representations in Computational and Theoretical Linguistics: the Potential for Mutual Enrichment[1]

Igor M. Boguslavsky
A. A. Kharkevich Institute for Information Transmission Problems, B. Karetnyj 15, Moscow, 103051, Moscow, Russia;
Universidad Politécnica de Madrid, 28040 Madrid, Spain
bogus@iitp.ru

Vyacheslav G. Dikonov
A. A. Kharkevich Institute for Information Transmission Problems
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
sdiconov@mail.ru

Evgeniya S. Inshakova
A. A. Kharkevich Institute for Information Transmission Problems
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
e.s.inshakova@gmail.com

Leonid L. Iomdin
A. A. Kharkevich Institute for Information Transmission Problems
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
iomdin@gmail.com

Alexandre V. Lazursky
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
lazursky@mail.ru

Ivan P. Rygaev
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
irygaev@jent.org

Svetlana P. Timoshenko
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
timoshenko@iitp.ru

Tatyana I. Frolova
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
tfrolova@gmail.com

Abstract

Research in semantics is actively conducted both in theoretical and computational linguistics, but the formulation of tasks, objectives and results of semantic research in the two communities are usually largely different. As a step towards reducing this gap and increasing the awareness of theoretical linguists about what computational linguists are doing, we examine meaning representation approaches in computational linguistics and contrast them with how this is done within one of the best-known theoretical approaches – the Meaning ⇔Text Theory.

Keywords: semantic representation, computational linguistics, theoretical linguistics, Meaning ⇔ Text Theory
DOI: 10.28995/2075-7182-2021-20-127-141

[1] Публикуется по специальному решению Редсовета

Семантические представления в компьютерной и теоретической лингвистике: потенциал взаимного обогащения

Богуславский И. М., Диконов В. Г., Иншакова Е. С., Иомдин Л. Л., Лазурский А. В., Рыгаев И. П., Тимошенко С. П., Фролова Т. И.

Аннотация

Семантические исследования активно ведутся как в теоретической, так и в компьютерной лингвистике, но постановки задач, цели и результаты таких исследований пересекаются слабо. В качестве шага, направленного на то, чтобы сократить этот разрыв и сделать более понятным для теоретических лингвистов то, что делают их коллеги в компьютерной лингвистике, мы рассматриваем некоторые способы представления значения предложения, принятые в компьютерной лингвистике, и сопоставляем их с тем, как представляются соответствующие явления в рамках одной из известных теоретических моделей – в модели "Смысл ⇔ Текст".

Ключевые слова: семантическое представление, компьютерная лингвистика, теоретическая лингвистика, модель «Смысл ⇔Текст»

1 Introductory remarks

From the start, the Dialogue conference has been a platform destined to bring together the two parts of our broad linguistic community – computational linguists (including mathematicians and engineers working in the field of computer natural language processing) and, so to speak, "linguistic" linguists engaged in theoretical and descriptive linguistics. Sadly, however, we have to state that the two communities have little understanding of, and little interest in, each other. If, for the sake of definiteness, we confine ourselves to semantics, we will see that even though semantics is actively developed by both communities, the tasks set by them, the objectives, and the results hardly ever come together. A mere look at the topics presented to the linguistic and computational linguistic sections of Dialogue – and to other similar forums – will suffice to conclude that the presenters speak different scientific languages and have little in common.

This cannot be accepted as normal: ultimately, we have one and the same object of research, the natural language. So, we should not simply acknowledge the difference in objectives, approaches and methods between computational linguistics (CL) and theoretical linguistics (TL) but, rather, endeavor to bridge this gap and raise the awareness of the “opposing sides” about these issues. Both communities will win if they get better acquainted with the practices and solutions of their colleagues from the other group.

The state-of-affair, as we see it, is that TL papers oriented at computational linguistics do appear from time to time, while there is virtually no movement in the opposite direction. We are unaware of any paper by CL workers targeted toward the TL community, attempting to compare the respective approaches. With this paper, we are making a step in this direction. It should be emphasized that the paper is primarily intended for non-CL linguists. Computational linguists will find no new results and no answers about the potential of the phenomena under discussion for the purposes of natural language processing. We hope, however, that the paper will offer something new to TL researchers. We are striving to compare the different approaches to an object which both communities consider as relating to them. We will talk to theoretical linguists about the objects that are relevant to them from the viewpoint which is close to them. Specifically, we will discuss the object very familiar to linguists interested in formal methods of linguistic description – the semantic representation of the sentence, as well as the information which this representation permits to express. Methods of semantic representations are of great interest to CL, too, because more and more applications require that text meaning be taken into account. Yet, as TL and CL have rather different objectives, whatever is interesting and important for one domain may be irrelevant for the other domain.

We will briefly present several types of semantic representations used in CL, comment on their similarities and differences and compare them with the approaches accepted in TL (exemplified by the Meaning ⇔ Text theory). A clear formulation of these similarities and differences seems important in order for theoretical linguists to better understand what are their neighbors engaged in, and to acquire a more stereoscopic view of their object of research. The readers interested in a more detailed overview of semantic representations made in CL perspective are referred to Abend, Rappaport 2017 and Bos, Abzianidze 2019.

2 Semantic Representation Requirements

We have to emphasize from the start that different research groups use different names for representations with which they work but we will call all of them s e m a n t i c r e p r e s e n t a t i o n s, or SemR, for consistency. The objective of any SemR is to reflect the meaning of a sentence but the amount of this reflection may vary depending on the particular purpose for which a given SemR is primarily intended. In particular, if SemR is to be used for inferences of any kind, it is highly desirable that it include certain logical information, e.g. negation and quantifiers, as well as some information on the lexical meaning, such as information on implicative predicates or presuppositions. For other purposes such information may be irrelevant and is not included into SemR.

This orientation to the purpose reflects certain important differences between CL and TL approaches. Normally, a theoretical linguist is not faced with the question why a phenomenon should be dealt with: if a phenomenon exists then it needs to be described. Contrariwise, CL takes into account two issues before tackling a phenomenon. First, it is considered to be of great importance whether the phenomenon is essential for applications. If application operation is not largely affected by the phenomenon, it is likely to be ignored rather than represented in SemR. Second, SemRs are often discussed in CL from the standpoint of corpus annotation, since the developers expect SemRs to be managed by machine learning techniques, which implies the need for a large corpus marked up with such representations.

In such discussions, the focus is often placed on what information, potentially useful for applications, can be quickly and uniformly marked up by annotators, rather than on what semantic information is conveyed by natural language sentences and therefore should be reflected in the SemR, as is customary for TL. Accordingly, the simplicity of SemR is often considered to be a vital advantage, because it affects the required level of annotators and the expected markup speed. To give an example, the Prague tectogrammatical corpus (Hajič 2002, Hajič et al. 2001) was developed by specially trained annotators, while semantic markup of the UCCA (Abend, Rappaport 2013a, b) and UDS corpora (White et al. 2017) was outsourced to much less skilled personnel. The GBM project took a middle stand as it used both qualified experts and unskilled annotators (Bos et al. 2017).

The volume of this paper does not allow us to review all types of SemR presented in the literature: there are quite a few of them, see e.g. AMR (Banarescu et al., 2013), Bridge (Bobrow et al. 2007), Compreno (Anisimovich et al. 2012), FrameNet (Baker et al. 1998), GMB (Bos et al., 2017), MRS (Copestake et al. 2005), OntoNotes (Hovy et al. 2006), PDC (Hajič 2002, Hajič et al. 2001), UNL (Uchida et al. 2005), OntoAgent (McShane, Nirenburg 2012), UCCA (Abend, Rappaport 2013a, b), UDS (White et al. 2017), SemETAP (Boguslavsky et al. 2020, Boguslavsky 2021), and de Salvo Braz et al. 2015. In section 4 below we will illustrate some representative approaches.

Before we proceed with the discussion, one important remark has to be made. Recently, an approach to semantics has gained popularity in CL, which avoids presenting the meaning of a linguistic object in the form of an explicit structure understandable by a human. This approach is based on the distributive hypothesis, which maintains that units occurring in similar contexts have similar meanings. In this approach, the meanings of words and even larger linguistic units are presented as vectors (ordered sequences of figures) built by analyzing the distribution of a given unit in a big collection of texts. Such vectors allow a quantitative assessment of the semantic proximity between the different units and are instrumental in the solution of certain other tasks (Lenci 2008), but they cannot help one obtain an exact idea of what a unit really means. Consequently, they give no clear answer to the essential questions of linguistic semantics: What does the word A mean? How does the meaning of A differ from the meaning of B? Therefore, vector methods of meaning representation cannot be considered transparent. As the aim of this paper is to discuss the compatible approaches to semantics in CL and TL, we will not touch on distributive models here.

3 Types of information represented in SemR

It is much more difficult to decide what information should be present in the semantic structure of a sentence than to answer a similar question about the syntactic structure. Indeed, in the case of syntactic structure, it is at least clear of what building blocks it should be made. The goal of the syntactic structure is to link **the words of the sentence**, which are observed directly. Of course, it is not always easy to construct an adequate structure but we know what units should be used. It is not at all obvious for the

semantic structure: what are the semantic elements that constitute a SemR of a sentence? How are they related to its words? How are these elements linked with each other? What information should appear in a SemR?

Below, we will comment on certain aspects essential in the comparison of different CL approaches to the construction of the SemR.

3.1 Which semantics is reflected by a SemR?

All SemRs referred to in Section 2 strive to abstract away from grammatical and syntactic idiosyncrasies inherent in natural languages. In particular, this is manifested in the fact that grammar words (auxiliary and support verbs, strongly governed prepositions and conjunctions, or articles) are removed from the sentence, passive constructions are replaced by active ones, nouns derived from verbs are reduced to the base verbs, etc. In many cases, such techniques help produce similar representations for different but synonymous constructions, and different representations for syntactically close but semantically diverging constructions. In this respect, many approaches view their semantic constructions as something close to deep syntactic constructions as understood in the Meaning ⇔ Text theory or the Prague school.

Such transformations are primarily confined to morphological and syntactic phenomena. The word semantics is often taken into account only partially: the words that are close in meaning receive the same representation or are reduced to one group of synonyms (WordNet synsets), or one frame of the FrameNet, or a different group of meanings. For example, the AMR approach builds one and the same SemR for sentences like *It may rain, It might rain, Rain is possible, It's possible that it will rain* (Banarescu *et al*. 2019). The attempts to represent the word meanings explicitly are very limited. AMR uses transparent lexical derivational models like *-able* or *-full*. The noun group *an edible sandwich* receives the same structure as *a sandwich that can be eaten.* To some extent, syntactic derivation is considered: *an attractive man* is represented in the same way as *a man who attracts.* In the structure used by Compreno, a word is not only referred to a class of the semantic hierarchy but may also be assigned a semantic feature, or semanteme, which explicates some component of the word meaning (Anisimovich et al. 2012).

A more detailed description of word semantics is represented in structures used in Bridge (Bobrow et al. 2007), OntoAgent (McShane, Nirenburg 2012) and SemETAP (Boguslavsky 2017, Boguslavsky et al. 2020). Among other things, Bridge takes account of implicative components of the word meaning, which specify the implications presumed by the word (as in *John managed to leave* ⟹ *John left*). OntoAgent and SemETAP decompose the word meaning into smaller semantic elements when necessary.

3.2 SemR nodes

There are two major approaches to the selection of units to be used as SemR nodes. Within the first approach the nodes are natural language words. In this case, the dictionary of a given natural language is frequently linked to a special lexical resource (which may have different names – ontology, dictionary of predicates, concepts, semantic classes or frames, WordNet) used to refer the words to a more general taxonomic category. For example, FrameNet refers the words *give*, *donate*, *gift* to the Giving frame, or concept. SemRs of the Bridge system refer each word to a set of WordNet synsets, in which at least one of the word's meanings is represented. Compreno associates the words with a specially designed hierarchy of semantic classes. Other versions of this approach, however, do not link the words as SemR units with any abstract conceptual entities. These are the cases of tectogrammatical structures of Prague Dependency Treebank (PDT) or Discourse Representation Structures of the GMB corpus.

Within the second approach, the nodes of SemR are elements of a semantic metalanguge (ontology). This can be exemplified by SemRs projects OntoAgent, SemETAP, and UNL. AMR structures occupy an intermediate position. The major part of semantic elements is composed of English words but there are several specially designed predicates, such as street-address, which has a list of arguments including house number, street, city, state, and zip code.

It is worth noting that the ontology (or another similar resource) plays different roles in these two approaches. In the first approach, the reference to a frame simply supplements the word in the SemR of the sentence but does not supersede it. Let us come back to the Giving frame above. The fact that it is referred to by the verbs *give, donate,* or *gift*, permits us to see and describe in a compact way the common features of the three verbs: they describe similar, though not identical situations and have the same sets of slots (frame elements). However, the semantic differences between the verbs remain unexplained. So

if a SemR contained the frame Giving instead of the verb *donate*, we would lose a part of the meaning because there is no full semantic identity between giving and donating. The approaches using FrameNet frames do not do that: the structure retains the initial verb (*donate*) beside the reference to the frame. In OntoAgent or SemETAP approaches, the ontology also contains concepts of the Giving type, but in the SemR such a concept will replace the verb *donate.* However, no meaning loss will occur, since the semantic representative of the verb *donate* is not the concept Giving alone but a certain construction composed of several concepts, which will explicitly express the semantic difference of *donate* from Giving.

3.3 Relations between SemR nodes

SemR elements are linked by relations. Of special importance are the relations between the predicates and their arguments. It is essential to show "who is doing what with which to whom", even though the boundaries between the argument (actant, or core) relations and non-argument ones (circumstantial, peripheral, or non-core) may be drawn variously.

All SemRs reflect verbal arguments. In many cases, arguments of some nouns and adjectives are present. We did not observe any SemRs (with the notable exception of SemETAP) that provided arguments of adverbs, despite the fact that such arguments are commonplace, see e.g. *far (from), independently (of), similarly (to), comparably (with), more (than), up (the hill)* etc. The relations between the predicates and their arguments are often marked with very general semantic roles, such as Agent, Theme, Patient etc. or using asemantic tags like ARG0, ARG1, ARG2 (as is common in PropBank or AMR). FrameNet takes a special stand, since many of the frame elements are specific for individual frames. For instance, the frame Arrest introduces the following specific argument relations (core frame elements): Authorities, Charges, Offence, Suspect.

To indicate non-argument semantic links between SemR elements, varied sets of semantic roles are used. One of the most popular sets of roles is proposed by the VerbNet project (Kipper et al. 2006).

3.4 Other types of information in SemR

In addition to the semantic relations between its elements, SemR may contain other types of data. We mentioned in Section 3.2 above that the SemR may contain a reference from a sentence word to a semantic element of a higher level of abstraction: a frame, a WordNet synset, or a class in the semantic hierarchy. Other sorts of data that can be marked in a SemR include information on anaphora or coreference (AMR, GMB, PDC, Compreno, OntoAgent, SemETAP, UNL). Such information may involve finding the antecedents of anaphoric pronouns (*When Mary woke up, she* [Mary or another person] *felt a sore throat*) or restoring the syntactically conditioned zero anaphora (*Having received* [Pete] *a bad mark, Pete decided to start* [Pete] *working* [Pete] *hard*).

In logically oriented SemRs, a logical structure is marked, which may include the negation, quantifiers and their scopes (GMB). The tectogrammatical structures of PDT are marked for thematic-rhematic articulation and the deep word order. In AMR structures, named entities are referred to the respective Wikipedia article:

```
(s / ship
  :wiki "RMS_Titanic"
  :name (n / name
        :op1 "Titanic"))
```

On the other hand, some SemRs are left with unmarked grammatical meanings, such as the number, tense, or definiteness/indefiniteness expressed by articles (AMR).

3.5 Reliance on a specific linguistic theory

Certain SemRs are built in accordance with a specific linguistic theory. So, PDT structures are oriented to Functional Generative Description, developed by the Prague School of Linguistics (Sgall, Hajičova and Panevová, 1986). Minimal Recursion Semantics (MRS) structures are closely connected with the Head-driven Phrase Structure Grammar (Pollard, Sag 1987). SemRs of GMB rely on the Discourse Representation Theory, or DRT (Kamp and Reyle 1993). SemETAP has been conceived within the

framework of the Meaning ⇔ Text theory (Mel'čuk 1974, 2012, 2013, 2014). By reasons of space, we cannot discuss this topic in more detail.

4 Examples of semantic structures

Having illustrated some of the general approaches to SemR construction we will now give a brief account of other approaches that are relatively rarely discussed.

4.1 Bridge (Bobrow et al. 2007)

The Bridge system, developed in Palo Alto Research Center (PARC), is designed to convert sentences into abstract knowledge representations (AKR). An AKR consists of three main parts: the conceptual structure, the contextual structure, and the temporal structure. The conceptual structure describes objects, their properties and the events in which they take part. The contextual structure relates this information to the real word, communicating whether the propositions mentioned in the conceptual structure exist in reality and reflecting the presuppositions of objects' existence. The temporal structure refers the time of the events mentioned to the moment of speech.

The content words are referred to the ontology, which in this case is WordNet. Every word receives references to all synsets in which they are present (if any).

The following example of a complete SemR was built for the sentence *John Smith discovered that three men had died.*

Conceptual Structure:

subconcept(discover:2, [detect-1, . . ., identify-5])
role(Theme, discover:2, ctx(die:5))
role(Agent, discover:2, Smith:1)
subconcept(Smith:1, [male-2])
alias(Smith:1, [John, Smith, John Smith])
role(cardinality restriction, Smith:1, sg)
subconcept(die:5, [die-1, die-2, . . ., die-11])
role(Theme, die:5, man:4)
subconcept(man:4, [man-1, . . ., world-8])
role(cardinality restriction, man:4, 3)

Contextual Structure:

context(t)
context(ctx(die:5))
top context(t)
context lifting relation(veridical, t, ctx(die:5))
context relation(t, ctx(die:5), crel(Theme, discover:2))
instantiable(Smith:1, t)
instantiable(discover:2, t)
instantiable(die:5, ctx(die:5))
instantiable(man:4, ctx(die:5))

Temporal Structure:

temporalRel(startsAfterEndingOf, Now, discover:2)
temporalRel(startsAfterEndingOf, Now, die:5)
Comments.

The conceptual structure introduces the concepts *discover*, *die* and *man* and specifies the WordNet synsets to which they belong. Smith is said to be male (this information is derived from knowledge that John is a name of a male).

The contextual structure contains two contexts: an upper level context **t**, which describes what the speaker communicates as a true statement about the world, and the internal context ctx(die:5), describing what John Smith has found – the fact that three men had died.

The verb *discover* has the presupposition that the subordinate predication is true. This component of the lexical meaning is described in the contextual structure with the relation called context lifting relation (veridical, t, ctx(die:5)). Thanks to this relation, an inference will be made that the predicate die:5 (instantiable(die:5, ctx(die:5)) is true.

The temporal structure reports that *discover* and *die* events took place prior to the moment of speech.

4.2 Tectogrammatical structure of PDT (Hajič 2002, Hajič et al. 2001)

The tectogrammatical level is the deepest level envisaged within the framework of Functional Generative Description, developed by the Prague linguistic school (Sgall, Hajičova and Panevová, 1986). This is the level of the semantic structure representation of a sentence, or, as the authors call it, the "linguistic meaning". In all, Functional Generative Description involves three levels of sentence representation. Beside the tectogrammatical level, there is an analytical level at which the surface syntactic structure is presented, and a morphological level at which the sentence is viewed as a sequence of lemmas supplied with morphological features. The Prague Dependency Treebank contains the structures of all three levels for every treebank sentence. We will discuss the deepest, tectogrammatical structure. Its major features are as follows.

- A tectogrammatical structure is a tree whose nodes are content words of a sentence. All grammatical words, including prepositions, conjunctions etc. are not represented, and the information conveyed by them is transferred to the attributes of content words.
- The nodes are linked by dependency relations, or functors (deep syntactic relations). There are five actant relations (Actor, Patient, Addressee, Origin, Effect), a large group of circumstantial relations and several technical relations marking the negation, coordination, apposition, foreign-language expressions etc. In all, 72 functors are used (Hajič 2002).
- The nodes are supplied with a set of attributes that convey varied information which enables one to synthesize the original sentence, or a sentence synonymous with it, from the tectogrammatical structure.
- The structure restores certain types of ellipsis and establishes certain types of coreference relations.
- The structure shows thematic/rhematic articulation, together with the so-called deep word order which reflects the position of a word in the old/new scale (the communicative dynamism).

Fig. 1 presents an example of a tectogrammatical structure.

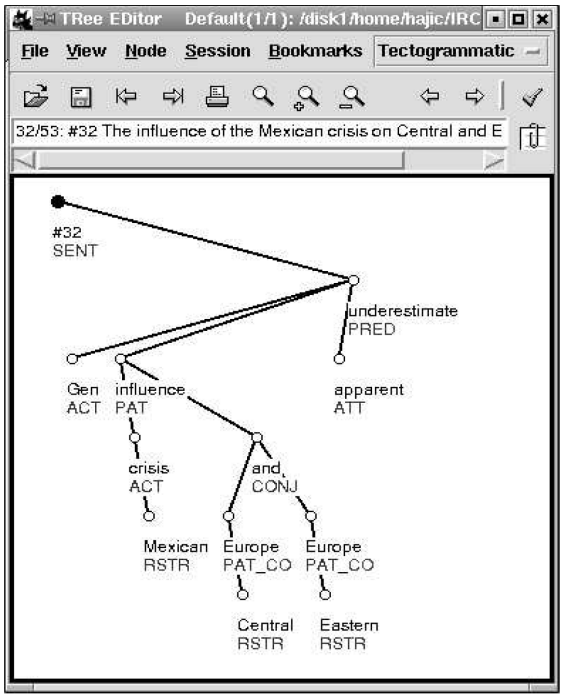

Fig. 1. The tectogrammatical structure for the sentence *The influence of the Mexican crisis on Central and Eastern Europe has apparently been underestimated* (Hajic et al. 2001)

4.3 DRS (Bos et al. 2017)

SemRs which constitute the Groningen Meaning Bank (GMB) semantic corpus are built on the basis of Discourse Representation Theory and are called Discourse Representation Structures (DRS). In contrast to most SemRs appearing in semantic corpora, DRS embraces many sorts of information. A DRS is a multilayer record that contains the predicate-argument structure, thematic roles, verbal tense, coreference, quantifier scopes, rhetorical relations, and presuppositions. Importantly, DRS are built for whole texts rather than individual sentences and reflect not only intrasentential but intersentential factors (coreference, rhetorical relations). Events are represented in the neo-Davidsonian style (Parsons 1990): every event is marked by an individual variable referring to this event.

For instance, the record **administer(e18)** denotes that the event **administer** has received the name **e18**, which represents this event in various propositions, e.g. **Agent(e18,x15)** – 'the agent of the event e18 is the entity x15' (see Fig.2).

A DRS consists of two parts. The upper part lists the entities participating in the situation being depicted, while the lower part represents their properties and interrelations. Fig.2 provides an example of SemR in the form of DRS.

x2 e18 x15 x19 t12 t20
named(x2, cayman_islands, org) administer(e18) Theme(e18, x2) named(x15, jamaica, loc) Agent(e18, x15) timex(x19,+1863XXXX) after(e18, x19) now(t12) e18 ⊆ t20 t20 < t12

Fig. 2. A DRS for the sentence *The Cayman Islands were administered by Jamaica after 1863* (Bos et al. 2017: 9).

Omitting some details, this DRS can be read as follows: «the event 'administer' has the named entity Jamaica as Agent and Cayman Islands as Theme. The event had place in the past, starting from 1863».

Words appearing in DRS are supplied with three classes of markup: named entities, such as Person, Location, Organization (in all, 7 varieties), indicators of Animacy degree, such as Human, Organization, Animal, Machine etc. (9 varieties), and WordNet synsets.

Semantic elements are linked with thematic roles borrowed from VerbNet. For certain kinds of expressions, such as two-noun compounds, possessive and temporal constructions, an implicit relation is generated in the form of a preposition. For example, the sentence *The Apple spokesman announced Wednesday that its new products will be released this week* contained 4 implicit relations: *the Apple spokesman* = '(spokesman) of Apple', *announced Wednesday = '*(announced) on Wednesday', *its (Apple) products = '*(products) by Apple', *released this week* = '(released) in this week' (Bos et al. 2017: 16-17). We see that the generated relations are not very semantic. Apparently, the expressions *announced Wednesday* and *released this week* contain the same semantic relation but as the relations are generated from words requiring different prepositions, the representations turn out different. Besides, the generated prepositions are normally polysemic, which is not accounted for in any way.

DRS can be directly translated into formulas of first order predicate logic, which allows one to use the available inference software.

4.4 Compreno (Anisimovich et al. 2012, Stepanova et al. 2016)

The integral syntactico-semantic structure of Compreno is a non-tree graph of dependencies, supplied with grammatical and semantic information. The nodes of the graph correspond to the words of the sentence and are connected by syntactic relations and semantic roles (Stepanova et al. 2016). Some types of ellipsis are restored. Each word is associated with some element of the semantic hierarchy. For

instance, the Russian word *бутерброд* is viewed as SANDWICH_AS_FOOD, and the word *говорить* as TO_SAY_SPEAK_TELL_TALK.

Besides the reference to a semantic or lexical class, the word may have a semantic feature (semanteme) which is a component of the word's lexical meaning. The semanteme facilitates the choice of a translational equivalent in a different language.

Fig. 3 gives an example of a Compreno SemR. Note that the SemR shows the restored subject *ученик* 'student' (one of the students ⇒ one student of the students').

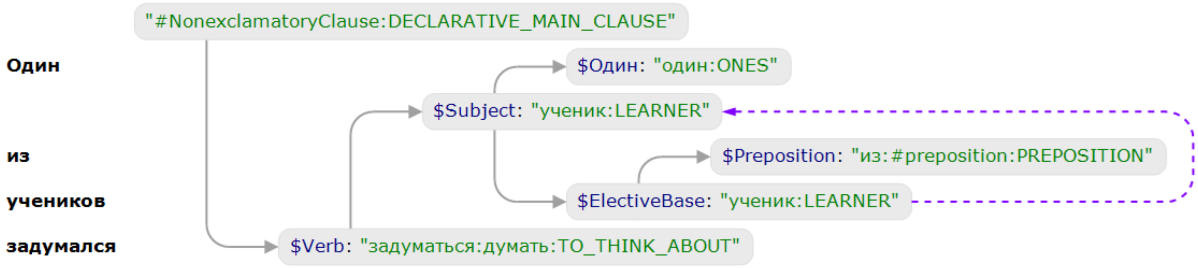

Fig. 3. Compreno SemR for the sentence
Один из учеников задумался 'One of the students started to ponder'

4.5 UNL (Uchida et al. 2005, Boguslavsky et al. 2005)

The Universal Networking Language (UNL) project was proposed by H. Uchida and developed by an international consortium for a number of years. The central idea is to develop a universal interlingual semantic representation (interlingua), which could serve as basis for the construction of a system of multilingual communication. Certain features of UNL are inherent in many SemR projects. In the amount of meaning, the semantic elements correspond to natural language lexemes. This means that on the one hand they match one lexical meaning of a word, rather than the whole vocable, while on the other hand they do not resort to decomposition of lexical meanings into smaller elements.

The semantic elements are linked to each other with dozens of binary relations resembling conventional semantic roles. The elements may be assigned additional features corresponding to modal and grammatical meanings. Semantic elements are organized into a hierarchy. A specific feature of UNL is a mechanism of dealing with lexical-semantic discrepancies between close though non-identical words of different languages. These discrepancies are described by the so-called constraints that are part of semantic elements. For example, semantic elements for the verb *marry* and its Russian counterpart *жениться* have specific constraints saying that the agent of the first one is a human (of either sex) and that of the second one is a male.

4.6 SemETAP (Boguslavsky et al. 2020, Boguslavsky 2021)

SemETAP is the semantic component of the functional model of language, ETAP-4, which implements the basic linguistic competences of humans – text understanding and text production. One of the key operations in text understanding is the extraction of all possible inferences. The model of understanding builds sequentially two semantic structures: the basic SemR and the enhanced SemR. The basic structure presents the direct meaning of the sentence, while the enhanced structure enriches it with a number of inferences which are construed on the basis of linguistic and extralinguistic knowledge accessible to the model. Both structures are built from the elements of a language-independent ontology, which thereby can be seen as a metalanguage of semantic description. All essential linguistic units are described in terms of ontological elements. The semantic description of many ontological concepts includes a decomposition of their meaning in the enhanced semantic structure into smaller components, which helps achieve a deeper text understanding, extract inferences and answer questions. For instance, the semantic structure of the concept 'envy' allows the model of understanding which processes a sentence like *Петя завидует Коле, что он нравится девушкам* 'Pete envies Nick because girls like him' and the question *Кто не нравится девушкам?* 'Whom don't girls like' provide the answer *Петя* 'Pete'.

Fig. 4 presents an example of a basic structure in SemETAP.

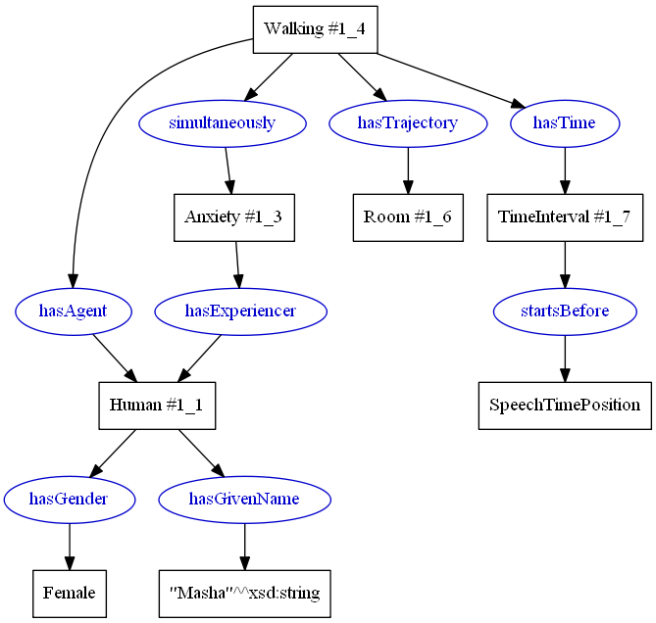

Fig. 4. The basic semantic structure for the sentence *Маша в волнении ходила по комнате* 'Masha walked about the room in anxiety'.

The structure can be read as follows: "A person of female gender, whose name is Masha, walks about the room and simultaneously experiences anxiety; this happened prior to the moment of speech".

5 Representation of Semantics in the Meaning ⇔ Text model

Of the whole range of theoretical semantic approaches, we chose the Meaning ⇔ Text model (MTT), developed by Igor A. Mel'čuk. It is a complete and very detailed theoretical model of language that describes how to represent a sentence at every level, including syntactic and semantic. This is why it is convenient to compare it with the semantic representations described above. Henceforth, we will call the semantic representation format of MTT — SemR-MTT for brevity, and semantic representations used in computational linguistics, SemR-CL.

We assume that the reader has sufficient knowledge of MTT and mention only two relevant publications: the first monograph describing the model (Mel'čuk 1974), and a comparatively recent three-volume work "Semantics: From meaning to text" (Mel'čuk 2012, 2013, 2015). We will present some characteristic features of MTT in comparison with SemRs-CL.

MTT is a strictly stratificational model. Every sentence receives formal representations at multiple levels: phonetic, morphologic, shallow-syntactic, deep-syntactic, semantic and conceptual. Most SemRs-CL are not viewed as parts of stratificational models and in many cases do not provide separate syntactic and semantic representations. Typical examples are AMR, which does not contain any dedicated representation of syntax, and Compreno, which has an integral syntactico-semantic structure. A notable exception is the tectogrammatical structure in PDT. It corresponds to the level of "linguistic meaning" (Hajič 2002) in the Functional Generative Description model and is contrasted with the morphological and analytical (surface-syntactic) structure. Among the levels of representation stipulated by MTT two may be seen as relevant counterparts of SemR-CL: deep-syntactic level (DSyntR-MTT) and semantic level (SemR-MTT).

The main purpose of (DSyntR-MTT) is to represent the syntactic structure of a sentence in a fashion that is abstracted away from any syntactic peculiarities of the sentence. Any syntactically conditioned grammemes, strongly governed prepositions and conjunctions, auxiliary verbs and other conventional words required by the grammar are eliminated from it. Syntactically synonymous constructions receive identical representation. The nodes of DSyntR-MTT are (almost) always content words, which may be supplemented with MTT lexical functions and a small number of special «fictitious» lexemes.

Most of the reviewed SemRs-CL are in many respects closer to the deep syntactic level of MTT than to its semantic level. First, the nodes of most SemRs-CL are content words of a natural language rather

than special semantic elements. Second, SemRs-CL closely reflect the syntactic skeleton of the source sentence. The best example here is AMR.

The types of relations between nodes of both DSyntR-MTT and SemR-CL are different from the relations between nodes of a purely syntactic (surface syntactic) structure, such as subject, direct object, indirect object, modifier, etc.

However, DSyntR-MTT uses universal deep-syntactic dependencies (numbered relations 1-7) to attach actants and **coordinative, quasi-coordinative, appenditive, attributive, descriptive-attributive relations. Semantic roles (agent, theme, experiencer, etc.) are not used. SemRs-CL use semantic roles in most cases. Some SemRs-CL (OntoNotes, AMR) reject the semantic roles to attach arguments to predicates and use indices instead: ARG0, ARG1, ARG2, etc.**

As noted above, DSyntR-MTT makes use of lexical functions in addition to content words. SemRs-CL completely ignore this large group of lexical means of expression, or, at best, omit support verbs corresponding to the lexical functions Oper/Func/Labor, as is the case in AMR. In the latter case, SemRs replace combinations like *make adjustments* with the single verb *adjust.* Meanwhile, dropping the support verb is not always possible. In particular, such omissions may produce anomalous results when a noun has a modifier. Compare *I had a feeling of relief – I felt relief* (the support verb *have* is dropped and the noun *feeling* is replaced with the corresponding verb) vs. *I had an unsettling feeling – *I felt unsettlingly* (the omission is not possible). This is one of the reasons why support verbs receive special treatment in MTT.

In certain SemRs-CL, including OntoAgent, SemETAP and UNL, the nodes may be special semantic elements instead of natural language words. Such structures bear more semblance to SemRs-MTT semantic structures and are farther away from the deep-syntactic level.

SemR-MTT is designed to represent the meaning of a source sentence irrespective of what words and syntactic structures were chosen to form it. The elements of SemR-MTT are not words but semantic elements — semantemes.

All content lexical units of the language receive definitions (semantic decompositions) consisting of semantemes. SemR-MTT can use different degrees of decomposition — from the minimal degree, when semantemes correspond to lexical senses of the corresponding words, to maximal degree, when decomposition reaches the level of semantic primitives. Minimal decomposition produces compact semantic structures, but in some cases it is not an adequate solution because certain semantic links cannot be revealed.

For example, the sentence *The green party won the majority at municipal elections* can be represented with a minimal decomposition of the semanteme ‘majority’. However, to adequately represent the sentence *The green party won a marginal majority at municipal elections* we need a deeper decomposition of ‘majority’. The modifier *marginal* does not refer to the whole meaning ‘majority’ (= ‘the number of votes «for» which is greater than the number of votes «against»’). Indeed, the number of votes secured by the greens can be very large. The modifier *marginal* concerns one of the inner semantic components of the word *majority*: the numbers of «for» votes is insignificantly greater than the number of «against» votes. It is impossible to show what the modifier’s contribution actually is without decomposing the meaning of ‘majority’ to the level, when its component ‘be greater than’ becomes explicit.

SemRs-CL almost never decompose lexical meanings. We know of only three projects in which the decomposition is performed in a substantial degree: SemETAP, OntoAgent, and Bridge.

Our comparison of DSyntR-MTT and SemR-MTT with SemRs-CL will not be complete without mentioning two “negative” facts. What features of semantic approaches used in computational linguistics are lacking in MTT?

First, MTT disregards all considerations of relevance of specific semantic components for computer applications. A semantic component is added to SemR-MTT of an expression if it is recognized as part of the meaning of that expression. Second, SemR-MTT never includes components that are produced by extralinguistic mechanisms: logical inference, world knowledge, common sense axioms, etc. All such phenomena are considered by MTT to belong to a level deeper than semantics, namely the level of conceptual representation. This distinction is accepted in computational linguistics too, but it is not drawn with the same rigor. There are approaches (DRT, OntoAgent and SemETAP) that make use of both linguistic and extralinguistic knowledge and aim to derive a wide range of logical inferences from texts.

6 Overview of Semantic Representation Features Considered

For the convenience of the readership, we summarize the main features of SemRs discussed above in Table 1. The cells for which we have no information are left blank. SemRs are classified by the following parameters:

- SemR nodes

[1] SemR nodes are NL words (+) or elements of a metalanguage (ontology, hierarchy of semantic classes, etc.) (-)
[2] If SemR nodes are NL words, they are linked to a higher level resource (WordNet, FrameNet, etc.)
[3] If SemR nodes are NL words, they are represented by a canonical variant (*may ⇒ possible, construction ⇒ construct*)

- SemR relations

[4] Predicate-argument relations are represented by semantic roles - Agent, Patient, etc. (+) and not by asemantic labels - ARG0, ARG1, etc. (-).

- Other information

[5] Anaphora/coreference is represented; ellipsis is restored.
[6] Information structure, Topic/Focus articulation is represented.
[7] Logical structure (quantifiers and their scope) is represented.
[8] Lexical meanings are decomposed (meaning postulates, lexical entailments, etc.)

- Levels of representation

[9] SemR is opposed to a syntactic structure.
[10] SemR is opposed to knowledge representation, which explicates different kinds of reasoning, e.g. logical entailment or common sense reasoning.

- Theory neutrality

[11] SemR is based on a specific linguistic theory.

- Existence of corpus

[12] There exists a corpus annotated by this type of SemR.
AMR: https://amr.isi.edu/
GMB: https://gmb.let.rug.nl/
PDC: https://ufal.mff.cuni.cz/pdt2.0/
UNL: http://www.unlweb.net/wiki/List_of_UNL_Corpora

	AMR	Bridge/PARC	GMB	PDC	Compreno	UNL	OntoAgent	Sem-ETAP	**MTT**
[1]	+	+	+	+	+	-	-	-	-
[2]	-	+	+	-	+	N/A	N/A	N/A	N/A
[3]	+	+		-	-	N/A	N/A	N/A	N/A
[4]	-	+	+	+	+	+	+	+	-
[5]	+		+	+	+	+	+	+	+
[6]	-			+	-	-	-	-	+
[7]	-		+	-	-	-		-	+
[8]	-	+	-	-	-	-	+	+	+
[9]	-	+	+	+	-	-	+	+	+
[10]	-	+	+	-	-	-	+	+	-
[11]	-	-	+	+	-	-	-	+	+
[12]	+	-	+	+	-	+	-	-	-

Table 1. Classification of SemRs

7 Conclusion

A wide range of SemRs are used in computational linguistics. Their main common feature is that they abstract away from grammatical and syntactic variety and try to represent different but synonymous constructions in similar fashion while contrasting syntactically similar constructions that differ in meaning. Another common feature is representation of predicate-argument relations, first of all for verbs, but

sometimes also for nouns and adjectives. Various sets of semantic roles are used to represent semantic relations between elements of SemR-CL. Many SemRs-CL also reflect other semantic phenomena: named entities (*Washington* as a human, city or US state), semantic derivates (*teacher* – 'someone who teaches'), coreference links, temporality, certain types of ellipsis.

Some SemRs-CL provide means to express presuppositions, rhetorical and discourse relations, scopes of quantifiers, implicit semantic relations. In rare cases of SemRs, partial decomposition of lexical meanings can be observed.

Often, the SemRs of sentences are produced on a large scale to create a semantic treebank. In this case the SemRs are mostly built by hand. A common way to facilitate this process is to ignore phenomena that are hard to tag for inexperienced annotators. Another method of building a semantic treebank is using an automatic semantic analyzer with subsequent manual editing of the output by experts (e.g. Bos et al. 2017).

Comparison of SemRs used in computational linguistics with the ways to express the meaning of language expressions in theoretical linguistics (illustrated by the Meaning ⇔ Text theory) showed that the two paths of semantic research have much in common but also reveal significant differences. As expected, the differences are mostly caused by different goal setting. Theoretical linguistics aims to embrace the full scope and complexity of linguistic phenomena and build explanatory models. Computational linguistics sees one of its primary goals in the creation of semantically annotated corpora (in order to be able to apply the full power of machine learning methods to work with SemRs).

If we agree that SemRs developed by theoretical linguists adequately represent the semantic phenomena encountered in the natural language, we should conclude that such SemRs can serve as a convenient source, from which the computational linguists can borrow methods of formal representation for certain new phenomena, as need be.

On the other hand, theoretical linguists can greatly benefit from the approaches to SemR construction developed in the field of CL. As noted before, most kinds of SemR-CL are embodied in text corpora annotated with such structures. Linguists have long learned to use annotated corpora in their research. The more diverse are the annotated data, the greater the value of the corpus for theoretical linguistics. A good example is the well-known FrameNet corpus, which provides vast information of ways in which semantic frames are expressed in English and some other languages. Another example is SynTagRus, which provides rich annotation for Russian texts, including morphological, syntactic, lexical-semantic, lexical-functional, anaphoric and some other types of annotation. Such corpora with deep annotation facilitate targeted search for data and statistical processing of the results.

We are sure that the wide range of semantic treebanks, already developed or being developed in CL, will be used fruitfully by theoretical and descriptive linguists to select, research and quantitatively assess the data.

Acknowledgments

This work was done with the financial support of a grant (No. 19-07-00842) from the Russian Foundation for Basic Research.

References

[1] Abend O., Rappoport A. UCCA: A semantic-based grammatical annotation scheme. // Proc. of IWCS. — 2013a. — p. 1–12.

[2] Abend O., Rappoport A. Universal Conceptual Cognitive Annotation (UCCA). // Proc. of ACL. — 2013b. — p. 228–238.

[3] Abend O., Rappoport A. The state of the art in semantic representation. // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Vancouver, Canada. Association for Computational Linguistics. — 2017. — p. 77–89.

[4] Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo. — 2012. — p. 90–103.

[5] Baker C. F., Fillmore, Ch., Lowe J. The Berkeley FrameNet project. // COLING-ACL '98. Proceedings of the Conference. — Montreal, Canada. — 1998.

[6] Banarescu L., Bonial C., Shu Cai, Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn Ph., Palmer M., Schneider N. Abstract meaning representation for sembanking. // Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. — Sofia, Bulgaria. — 2013. — p. 178–186.
[7] Banarescu L., Bonial C., Shu Cai, Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn Ph., Palmer M., Schneider N. Abstract Meaning Representation (AMR) 1.2.6 Specification. — 2019. — https://github.com/amrisi/amr-guidelines/blob/master/amr.md#special-frames-for-roles
[8] Bobrow D., Cheslow B., Condoravdi C., Karttunen L., Holloway King T., Nairn R., de Paiva V., Price Ch., Zaenen A. PARC's Bridge and Question Answering System. // Proceedings of the GEAF 2007 Workshop Tracy Holloway King and Emily M. Bender (Editors). CSLI Studies in Computational Linguistics. Ann Copestake (Series Editor). — 2007.
[9] Boguslavsky I. Semantic Descriptions for a Text Understanding System. Computational Linguistics and Intellectual Technologies. // Proceedings of the International Conference "Dialog" [Komp'yuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]. — 2017. — p. 14–28.
[10] Boguslavsky I. Semantic analysis supported by inference in a functional model of language [Semanticheskij analiz s oporoj na umozakljuchenija v funktsional'noj modeli jazyka]. // Problems of linguistics [Voprosy jazykoznanija], № 1 — 2021. — pp. 29-56.
[11] Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L. The UNL Initiative: An Overview. Lecture Notes in Computer Science. // Volume 3406, Springer Berlin / Heidelberg. — 2005. — p. 377 – 387.
[12] Boguslavsky I.M., Dikonov V.G., Frolova T.I., Iomdin L.L., Lazursky A.V., Rygaev I.P., Timoshenko S.P. Full-fledged Semantic Analysis as a Tool for Resolving Triangle-Copa Social Scenarios. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'yuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], issue. 19 (26), — 2020 — p. 106-118.
[13] Bos J., Abzianidze L. Thirty Musts for Meaning Banking. // Proceedings of the First International Workshop on Designing Meaning Representations. — Florence, Italy, August 1st. — 2019. — p. 15–27.
[14] Bos J., Basile V., Evang K., Venhuizen N.J., Bjerva J. The Groningen Meaning Bank. // Ide N., Pustejovsky J. (eds). Handbook of Linguistic Annotation. — Springer, Dordrecht. — 2017.
[15] Copestake A., Flickinger D., Pollard C., Sag I. Minimal recursion semantics: An introduction. Research on Language and Computation 3:281–332. — 2005.
[16] Hajič J. Tectogrammatical Representation: Towards a Minimal Transfer In Machine Translation. // Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks. Association for Computational Linguistics. — 2002. — p. 216–226.
[17] Hajič J., Hladká B., Pajas P. The Prague Dependency Treebank: Annotation Structure and Support. // IRCS Workshop on Linguistic Databases. — 2001. — p. 105--114.
[18] Homola P. Neo-Davidsonian Semantics in Lexicalized Grammars. // Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013). — 2013. — p.134-140.
[19] Hovy E., Marcus M., Palmer M., Ramshaw L., Weischedel R. OntoNotes: the 90% solution. // Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. — Stroudsburg, PA, USA. — 2006. — p. 57–60.
[20] Kamp H., Reyle U. From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT. — Kluwer, Dordrecht. — 1993.
[21] Kipper K., Korhonen A, Ryant N., Palmer M. Extending verbnet with novel verb classes. // Proceedings of LREC, volume 2006. — 2006. — p. 1
[22] Lenci A. Distributional semantics in linguistic and cognitive research. // Special issue of the Italian Journal of Linguistics, Rivista di Linguistica 20.1. — 2008. — p. 1-31.
[23] McShane M., Nirenburg S. A knowledge representation language for natural language processing, simulation and reasoning. // International Journal of Semantic Computing Vol. 6, No. 1, 3_23. — 2012.
[24] Mel'čuk I.A. An essay of the theory of linguistic "Meaning ⇔ Text"models. Semantics. Syntax. [Opyt teorii lingvističeskix modelej "Smysl ⇔ Tekst". Semantika, Sintaksis.]. — Science [Nauka], Moscow. — 1974.
[25] Mel'čuk I. Semantics*: From Meaning to Text.* Vol. 1. — Amsterdam/Philadelphia: John Benjamins. — 2012.
[26] Mel'čuk I. Semantics*: From Meaning to Text.* Vol. 2. — Amsterdam/Philadelphia: John Benjamins. — 2013.
[27] Mel'čuk I. Semantics*: From Meaning to Text.* Vol. 3. — Amsterdam/Philadelphia: John Benjamins. — 2015.
[28] Parsons T. Events in the semantics of English: A study in subatomic semantics. — Cambridge, MA: The MIT Press. — 1990.
[29] Pollard, C., Sag I. Information-based syntax and semantics. Volume 1. Fundamentals. // CLSI Lecture Notes 13. — 1987.
[30] de Salvo Braz R., Girju R., Punyakanok V., Roth D., Sammons, M. Knowledge Representation for Semantic Entailment and Question-Answering. // IJCAI'05: Workshop on Knowledge and Reasoning for Question Answering. — 2005.

[31] Sgall, P., Hajičová E., Panevová J. The Meaning of a Sentence in its Semantic and Pragmatic Aspects. — Prague - Amsterdam: Academia – North-Holland. — 1986.
[32] Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., Skorinkin D. A. Information Extraction Based on Deep Syntactic-Semantic Analysis. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016” Moscow, June 1–4. — 2016.
[33] Uchida H., Zhu M., Della Senta T. Universal Networking Language. Edition 2. — UNDL Foundation. Geneva. — 2005.
[34] White A., Reisinger D., Sakaguchi K., Vieira T., Sheng Zhang, Rudinger R., Rawlins K., Van Durme B. Universal decompositional semantics on universal dependencies. // Proc. of EMNLP. — 2016. — pp. 1713–1723.