Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Constructing a Semantic Corpus for Russian: SemOntoCor

Igor M. Boguslavsky
A. A. Kharkevich Institute for Information Transmission Problems,
Moscow, Russia;
Universidad Politécnica de Madrid,
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

The SemOntoCor project focuses on creating a semantic corpus of Russian based on linguistic and ontological resources. It is a satellite project with regard to a semantic parser (SemETAP) being developed, the latter aiming at producing semantic structures and drawing various types of inferences. SemETAP is used to annotate SemOntoCor in a semi-automatic mode, whereupon SemOntoCor, when reaching sufficient maturity, will help create new parsers and other semantic applications. SemOntoCor can be viewed as a further step in the development of SynTagRus with its several layers of annotation. SemOntoCor builds on top of the morpho-syntactic annotation of SynTagRus and assigns each sentence a Basic Semantic Structure (BSemS). BSemS represents the direct layer of meaning of the sentence in terms of ontological concepts and semantic relations between them. It abstracts away from lexico-syntactic variation and in many cases decomposes lexical meanings into smaller elements. The first phase of SemOntoCor consists in annotating a Russian translation of the novel "The Little Prince" by Antoine de Saint-Exupery (1532 sentences, 13120 tokens).

Keywords: semantic corpus; semantic parser; ontology; Saint-Exupery
DOI: 10.28995/2075-7182-2023-22-12-25

Разработка семантического корпуса русского языка: SemOntoCor

Богуславский И. М.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия;
Universidad Politécnica de Madrid,
Мадрид, Испания
bogus@iitp.ru

Диконов В. Г.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
sdiconov@mail.ru

Иншакова Е. С.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
e.s.inshakova@gmail.com

Иомдин Л. Л.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
iomdin@gmail.com

Лазурский А. В.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
lazursky@mail.ru

Рыгаев И. П.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
irygaev@jent.org

Тимошенко С. П.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
timoshenko@iitp.ru

Фролова Т. И.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
tfrolova@gmail.com

Аннотация

Проект SemOntoCor ставит своей целью создание семантического корпуса русского языка на основе лингвистических и онтологических ресурсов. Этот проект развивается параллельно с разработкой семантического анализатора (SemETAP), нацеленного на построение семантических структур предложения и извлечение из них разного рода следствий. SemETAP используется для разметки SemOntoCor в полуавтоматическом режиме. С другой стороны, после того, как SemOntoCor достигнет достаточной зрелости, он сможет использоваться для разработки новых семантических анализаторов и для других семантических задач. SemOntoCor можно рассматривать как следующий шаг в развитии синтаксического корпуса SynTagRus, имеющего несколько уровней разметки. При разметке SemOntoCor на вход поступает морфо-синтаксическая разметка в формате SynTagRus, а на выходе строится базовая семантическая структура (BSemS). Эта структура представляет непосредственное значение предложения в терминах онтологических концептов, соединенных семантическими отношениями. Она абстрагируется от лексико-синтаксического многообразия естественного языка и во многих случаях осуществляет разложение лексического значения на более мелкие компоненты. Первая очередь SemOntoCor представляет собой разметку русского перевода повести-сказки Антуана де Сент-Экзюпери «Маленький принц» (1532 предложения, 13120 токенов).

Ключевые слова: семантический корпус; семантический парсер; онтология; Сент-Экзюпери

1 Introduction

Among various semantically annotated corpora, few combine multiple levels of annotation into one formalism. A well-known example is OntoNotes [Hovy et al. 2006] which is a resource comprising syntax, predicate-argument structure, word senses and co-reference. Another example is the Groningen Meaning Bank [Bos et al. 2017] that aims at integrating various linguistic phenomena, including predicate-argument structure, scope, tense, thematic roles, rhetorical relations and presuppositions within the formalism of the Discourse Representation Theory. On the other hand, an obvious fact is that most semantically annotated corpora that exist nowadays concentrate on English. This language is supported by many corpora built within different frameworks and annotated according to different annotation schemes. One of the rare exceptions that stands out in various respects is the Prague Dependency Corpus, which contains deeply annotated Czech texts (https://ufal.mff.cuni.cz/pdt3.0/data).

As far as Russian is concerned, it has a remarkable Russian National Corpus (RNC) which includes an extensive set of subcorpora, which cover over 2 billion words (ruscorpora.ru). Most of the subcorpora are annotated with morphological tags, and the main subcorpora (general subcorpus, newspaper subcorpus and a few others) have been recently supplied with syntactic features (including universal dependency features) and lexical semantic tags. Most of these annotation types have been produced automatically and have not been checked by experts on a mass scale. The SynTagRus subcorpus of RNC stands out as it contains several types of deep annotation, carried out in a uniform formalism according to the same theoretical framework. Namely, it is annotated with morphological features (including POS tags), dependency syntactic structures, word senses, anaphoric links, lexical functions (in terms of the Meaning-Text theory by Mel'čuk [1974]), micro-syntactic constructions, ellipsis, and temporal links. An important feature of these annotations is that, although many of them were carried out by means of software tools, all were thoroughly manually revised by the experts. Introducing a deeper level of annotation and aligning SynTagRus texts with semantic structures is a natural step further. In this paper, we describe an ongoing project aiming at performing this step. We began to compile a corpus annotated with what we call Basic Semantic Structures (BSemS). These structures are built on top of the existing morpho-syntactic annotation of SynTagRus and thus constitute the next higher level of sentence representation. It is important to note that our goal is not to annotate certain phenomena in a linguistically isolated way but to integrate all relevant semantic phenomena in a unified representation.

The structure of the paper is as follows. Section 2 outlines related work. Section 3 explains the framework of the project. Basic Semantic Structures we are constructing to annotate the corpus constitute an intermediate level of representation adopted in our semantic model. Section 4 presents BSemS in finer detail and explains some of its salient features. Section 5 describes the format of the corpus and shows how it is annotated by means of a special tool developed to facilitate the mark-up. Section 6 concludes the paper.

2 Related work

Existing semantic corpora may be classed into several groups by the types of information they provide.

1) The first group embraces semantically tagged corpora which annotate texts with word senses, ontology concepts or abstract semantic descriptors but do not provide information about any semantic relations. Some of them cover all non-grammatical words, others - only specific classes or words. Such corpora are used to test and train word-sense disambiguation tools. Some examples are:

- Semcor [Fellbaum et al, 1998] tags words with Wordnet synset references.
- Russian National Corpus [Raxilina et al, 2009, Kustova et al, 2005] contains automatically produced facet semantic tagging with semantic descriptors.
- Colorado Richly Annotated Full-Text (CRAFT) Corpus [Bada et al, 2012] is a collection of 97 full-length, open-access biomedical journal articles annotated with concepts from 9 different medical ontologies in parallel.

An overview of other resources of the same kind was presented in the paper «A Survey of WordNet Annotated Corpora» [Petrolito, Bond, 2014].

2) The second group consists of the corpora that explicate semantic relations between words or senses/concepts that replace the words. They provide some kind of semantic structures, which may be built in accordance with a certain linguistic theory or be theory-neutral. There are several theories

defining the representation for the sentence or whole text meaning that became the basis for corpus development projects. An overview of various semantic representations, including Abstract Meaning Representation (AMR), Discourse Representation Structures (DRS), Universal Networking Language (UNL), Tectogrammatical Representation (PDT) and more can be found in [Boguslavsky et al, 2021].

The most popular approach within this group is Frame Semantics [Fillmore, 1976]. There are multiple corpora that annotate propositional (predicate-argument "Who did what to whom?") structures within sentences using FrameNet [Baker, Fillmore et al., 1998, Ruppenhofer et al. 2007, 2016] or Verbnet [Kipper et al., 2008] dictionaries. These dictionaries serve as repositories of frames – prototypical situations that include a particular verbal sense (predicate), all necessary participants of the situation (arguments) and the roles they play. Such corpora are commonly called proposition banks or propbanks. They take different approaches towards the description of the roles. FrameNet has a very specific representation while the original PropBank corpus has the most general representation. The Verbnet approach takes the middle stand between the two. For example, given the ingestion sense of the verb *eat* in the sentence "Cynthia ate the peach with a fork", the respective representations for each would be:

- FrameNet: Cynthia(*ingestor*) ate(*predicate*) the peach(*ingestible*) with a fork(*instrument*).
- VerbNet: Cynthia(*agent*) ate(*predicate*) the peach(*patient*) with a fork(*instrument*).
- PropBank: Cynthia(*arg0*) ate(*predicate*) the peach(*arg1*) with a fork(*argm-manner*).

Some examples of proposition banks are:

- The original English Proposition Bank (PropBank) [Palmer et al, 2005, Palmer 2002] and a whole family of Propbanks in other languages than English.
- A similar resource called Nombank [Meyers et al, 2004] annotates predicative nouns.
- FrameNet Corpus [Bauer et al, 2012] contains parser-generated dependency structures (with POS tags and lemmas) for all FrameNet 1.5 sentences, with nodes automatically associated with FrameNet annotations.
- OntoNotes [Hovy et al, 2006, Weischedel et al, 2009] is a large multilingual corpus. The annotation includes parse trees, predicate argument structures (PropBank/NomBank style), word senses linked to an ontology, coreference, and named entities. The languages covered are English, Chinese, and Arabic with a significant amount of parallel data.
- FrameNet-Annotated Textual Entailment (FATE) [Burchardt, Pennacchiotti, 2008] is a manually crafted corpus for Recognizing Textual Entailment (RTE) tasks with FrameNet annotation. It features a new annotation schema based on full-text annotation of so-called relevant frame evoking elements. FATE annotates frames unknown in the FrameNet, anaphoric expressions in frames and constructions important for RTE, including support and copula verbs, existential constructions, modal expressions, metaphors.

3 SemETAP semantic model

The SemOntoCor corpus is a collection of Russian texts annotated with BSemSs built in accordance with the SemETAP semantic model [Boguslavsky 2017, Boguslavsky et al. 2018, Boguslavsky et al. 2019]. In its turn, this model is a component of a general-purpose rule-based linguistic processor ETAP-4, which implements basic linguistic competences of humans – text understanding and text production [Apresian et al. 2003]. ETAP-4 is built within the framework of the Meaning – Text Theory by I. Mel'čuk [1974, 2012, 2013, 2015]. SemETAP reuses the non-semantic modules of ETAP-4 – the morphological analyzer, the syntactic dependency parser, and the normalization submodule. SemETAP is used for annotating SemOntoCor in a semi-automatic mode (see Section 4 for details).

Our approach to semantics is in many respects similar to that of [McShane, Nirenburg 2021], although many linguistic and methodological solutions are different. We proceed from the assumption that the depth of understanding is growing with the number and sophistication of inferences we can draw from the text. In order to obtain inferences, we make intensive use of both linguistic and background knowledge. The former is incorporated in the dictionary and the grammar, and the latter is stored in the ontology. In many cases, explicit decomposition of words and ontology concepts is used to produce additional inferences and thus achieve a deeper understanding.

We distinguish two levels of our semantic structures. Basic semantic structure (BSemS) presents the direct meaning of the sentence, while Enhanced semantic structure (EnSemS) extends BSemS by means of a series of inferences construed on the basis of linguistic and extralinguistic knowledge accessible to the model. The model produces both reliable inferences (*John forgot to take the pill ⇒ John did not take the pill*) and plausible expectations (*John went to Paris at moment t1 ⇒ John is in Paris at moment t2>t1*).

Both structure types (BSemS and EnSemS) are built from the elements of a language-independent ontology, OntoETAP, which thereby can be seen as a metalanguage of the semantic description. The ontological elements (concepts and individuals) have different kinds of properties in OntoETAP, such as class/subclass, class/individual, semantic slots a concept can take, etc.

From the formal point of view, semantic structures of both BSemS and EnSemS are Directed Acyclic Graphs (DAG) with individuals at the nodes and arrows labeled with semantic relations. They are represented in OWL and written in the RDF format, i.e. as sets of triples of the type *relation (Ontoelement-1, Ontoelement-2)*, where *relation* is an object or data property of the ontology, and *Ontoelement-i* is a variable or a constant denoting an individual. The RDF formalism was chosen because, on the one hand, it is very flexible and expressive, and on the other hand, it is supported by a wide range of tools and is easily integrated with many Semantic Web applications.

It should be noted that SemOntoCor, like SynTagRus and some other corpus projects, has double identity. On the one hand, it can be perceived as an autonomous semantic resource, and on the other hand it forms a unified complex with the SemETAP semantic parser, ontology and the inference engine. Below, we will concentrate on BSemS, because it is this type of structure that constitutes our corpus.

4 Basic Semantic Structures

When constructing a semantic representation of a natural language text, one of the most essential requirements is its ability to abstract away from formal and syntactic variation, namely to assign similar structures to different constructions that have a similar meaning, and to assign different structures to constructions that have different meanings, despite their surface similarity [Abend, Rappaport 2017].

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. The most important type of information that semantic structures seek to convey is the predicate-argument skeleton of the sentence, that is, the information about "who is doing what to whom". Here is the BSemS of one of the SemOntoCor sentences as an example:

(1) *Narisuj mne barashka...* ‘Draw me a lamb’

```
(2) Drawing
        hasAgent UtteranceAddressee
        hasObject Sheep
                hasGender Male
                isObjectOf HavingSize
                        hasDegree LowDegree
        hasBeneficiary UtteranceSpeaker
        isTopicOf Urging
                hasAgent UtteranceSpeaker
                hasRecipient UtteranceAddressee
                hasTime SpeechTimePosition
```

BSemS (2) representing sentence (1) can be read as follows: «The speaker verbally encourages the addressee to draw a small male sheep for the speaker; the time of encouraging is the time of speech».

For all the simplicity of sentence (1), the structure (2) allows us to see how SemOntoCor addresses some of the main problems facing semantic corpora. These are the representation of word meanings (Subsection 4.1), the representation of relations between words (Subsection 4.2) and the representation of grammatical meanings (Subsection 4.3).

4.1 Word meaning

The semantic units (nodes) that form BSemS are not natural language words (in our case, Russian), but elements of the OntoETAP ontology. This makes BSemS largely language-neutral. In semantic corpora, there is often a one-to-one correspondence between full-fledged words in a sentence and semantic elements (with the precision up to synonymy) (see [Boguslavsky et al. 2021, section 3.2] for more details). In our example (1), such a correspondence exists for two words in the sentence – *narisuj* (Draw) and *mne* (UtteranceSpeaker). The third word in the sentence – *barashek* 'lamb' - is represented by multiple nodes of the structure simultaneously (Sheep hasGender Male isObjectOf (HavingSize hasDegree LowDegree) – 'a small male sheep'). Partial decomposition of the lexical meaning with a group of several semantic elements is widely used in SemOntoCor.

This approach has both advantages and disadvantages. One advantages is that it allows producing similar representations for different synonymous expressions: *zapel* ('began-to-sing') – *nachal pet'* ('began to sing') = Begin hasObject Singing, *ispugal ee* ('frightened her') – *zastavil ee bojat'sja* ('made her fear') = Cause hasObject Fear. Even if the expressions are not synonymous but have significant semantics in common, the decomposition makes it possible to make explicit both the common components and the differences. For instance, in the AMR corpora (https://github.com/amrisi/amr-guidelines/blob/master/amr.md), the name of a feature scale and a specific range of that scale (such as *age – old/young, weight – heavy/light, price – expensive/cheap*) are expressed as separate concepts that are not related to each other. In SemOntoCor, the links between such meanings are presented in a clear and graphic way: *age* – HavingAge, *old* – (HavingAge hasDegree HighDegree), *young* – (HavingAge hasDegree LowDegree).

Decomposition allows avoiding uncontrolled introduction of a large number of conceptually similar concepts. For example, the Russian word *tigr* 'tiger' denotes any animal of the given species, while the word *tigritsa* 'tigress' denotes only a female of such an animal. The principle of mutual one-to-one correspondence between words and concepts requires the introduction of two separate concepts – Tiger and FemaleTiger – in the ontology. Decomposition of the lexical meaning allows us to use only one concept, defining a female tiger as (Tiger hasGender Female). Similarly, *frantsuz* 'French person' is interpreted as the construction (Human hasNationality France) (= a person who is a French citizen), *frantsuzhenka* 'French woman'- as (Human hasNationality France hasGender Female), *parizhanin* 'Parisian' – as (Human livesIn Paris) (= a person living in Paris). Decomposition can be used to distinguish and simultaneously capture the similarity of social roles (for example, *monarx1* = MonarchRole) and people who perform social roles (*monarx2* = (Human hasSocialRole MonarchRole)).

The downside of the lexical meaning decomposition approach is that many words refer to complex phenomena, and decomposing them into smaller components would result in very unwieldy structures. To avoid this, we adopted a compromise approach in BSemS, where only words that allow for a small number of components, like the examples above, are decomposed.

The general principle is that the full-valued words are matched with concepts that are synonymous or quasi-synonymous with them. At the same time, we try not to clutter the ontology with different concepts that are similar in meaning. Therefore, when the meaning of a word can be reduced to a basic concept, supplemented by a small number of refinements, we resort to decomposition. As for more complicated decompositions, we postpone them to the level of EnSemS, where different kinds of inferences are performed, including those based on common sense. For more on EnSemS, the reader is referred to [Boguslavsky 2017, Boguslavsky et al. 2018, Boguslavsky et al. 2019].

To give the reader a better understanding of the non-isomorphism between the words of the sentence and their representation in BSemS, we provide a couple more examples.

One systematic example is the reduction of different words to the same concept through lexico-syntactic derivation. Many words refer to the same situation from different angles: *John is married to Ann – Ann is John's wife – John is Ann's husband.* Obviously, the words used in these sentences are not synonyms, but their semantic representations should make it clear that they denote the same situation. In SemOntoCor these words are assigned the same concept (in this case – Spouse), and the difference between them is accounted for by different relations between the concepts. In this example, we were dealing with nominal lexico-syntactic derivatives of the verb – *husband* and *wife.* There exist also adverbial and adjectival derivatives, that are also reduced to the main predicate. Cf. (3) and (4).

(3) *Ja dumaju, chto pojdet dozhd'* 'I think it is going to rain'

Believe
    hasExperiencer UtteranceSpeaker
    hasObject Raining

(4) *Po-moemu, pojdet dozhd'* 'in my opinion, it is going to rain'

Raining
    isObjectOf Believe
        hasExperiencer UtteranceSpeaker

Another example of the same type involves adjectives that refer to the same concept but characterize its different arguments. For example, *ispugannyj* 'frightened' is a property of the one who if afraid of something, while *strashnyj* 'scary' characterizes something that causes fear. In SemOntoCor, such adjectives are represented by the same concept but are connected to the concept they modify by different semantic relations:

ispugannyj malchik 'frightened boy' – Boy isExperiencerOf Fear
strashnaja situacija 'scary situation' – Situation isObjectOf Fear

Yet another example of non-isomorphism between the text and its BSemS are titles. A phrase like *roman Remarka "Tri tovarishcha"* meaning 'the novel by Remarque "Three Comrades"' is rendered by a structure that contains both the original title and its meaning:

Novel
    hasName "Tri tovarishcha"
    hasNameMeaning Friend
        hasQuantity 3

A frequent source of nodes that have no direct correspondence in the text is omission of arguments. For example, in the phrase *Petya xochet poprosit' Kolyu ujti* meaning 'Petya wants to ask Kolya to leave', the BSemS structure makes explicit the omitted subjects of the infinitives: *poprosit' – Petya* meaning 'ask – Petya', and *ujti – Kolya* meaning 'leave – Kolya'.

4.2 Relations between the nodes in BSemS

As mentioned above, BSemS is a graph consisting of OntoETAP ontology elements in its nodes and semantic relations as its edges. Several dozen relations are used. The most frequent ones are: hasAgent, hasObject, hasObject2, hasExperiencer, hasTime, hasLocation, hasDegree, hasQuantity, hasAttribute, etc[1].

Each relation can have an inverted variant, such as hasAgent – isAgentOf, hasObject – isObjectOf. Inverted relations allow expressing the difference in communicative dominance. More information on this can be found in [Mel'čuk 2015: 311-324]. Cf. *Malchik bezhit* 'the boy runs' – (Running hasAgent Boy). *Malchik, kotoryj bezhit (begushchij malchik)* 'the boy who runs (running boy)'– (Boy isAgentOf Running). Cf. also examples (3) and (4) above.

As far as the propositional content is concerned, the structures (Running hasAgent Boy) and (Boy isAgentOf Running) are completely equivalent. Therefore, for tasks that involve only propositional content, communicative dominance is irrelevant and inverted relations can be safely replaced with non-inverted ones. However, for other tasks, information on communicative dominance may be valuable. Structures that differ in communicative dominance are used in different contexts and are built into discourse in different ways. Cf. *Malchik bystro bezhal* meaning 'the boy was running fast' and *Skorost' bega malchika byla vysokoj* meaning 'the boy's running speed was high'. Obviously, the communicative

[1] A complete commented list of relations used for annotation can be found at https://docs.google.com/document/d/1W469sCt-ne7DB1yS3QM_hzpCJM_yuhJp

dominance is crucial for any task dealing with text generation and discourse cohesion. It is also important that the use of inverted relations makes BSemS easier to understand as it aligns more directly with the syntactic structure (cf. structures (3) and (4) above).

Besides inverting, each relation can be reified, i.e. it can come in the form of a concept. If additional information, such as time or modality, needs to be conveyed along with the relation, one cannot use the relation and should use a reified concept. For example, the meaning of localization in the noun phrase *dom v lesu* 'a house in the forest' can be conveyed by the relation: (House hasLocation Forest). However, if we need to characterize this situation as having taken place in the past, the relation cannot be used and we should use a concept (Location) instead:

```
Dom naxodilsja v lesu
Location
        hasObject House
        hasObject2 Forest
        hasTime TimeInterval
                before SpeechTimePosition
```

4.3 Uniform representation of lexical and grammatical meaning

For representing grammatical meanings, BSemS does not use grammatical labels such as present, past, imperative, interrogative, etc. Instead, we apply the same concepts that represent lexical meanings. In structure (2) above, the imperative form of the verb *narisovat'* 'to draw' is conveyed by the semantic structure meaning 'the speaker encourages the addressee to draw' by means of the concept Urging. This concept is also used to represent lexical (and not grammatical) markers of encouragement, for example *Peter encouraged (urged) Bill to stay.*

Similarly, to represent grammatical interrogation (*When will you come?*) we apply the concept Questioning, which is also used to represent such words as *ask, question, enquire, interrogate* etc. Thanks to the uniform representation of grammatical and lexical meanings, sentences of different structures obtain similar (or identical) BSemS. For example, sentences *Kogda ty pridesh'?* 'When will you come?' and *Ja sprashivaju tebja, kogda ty pridesh'* 'I am asking you when you will come' correspond to the same BSemS. Besides, there is also a natural opportunity to establish co-referentiality between grammatical and lexical expressions, such as: *Ty pridesh' zavtra? Etot vopros byl neumestnym.* 'Will you come tomorrow? The question was inappropriate'. If grammatical interrogation were conveyed by a grammatical label, it would be hard to infer that it is co-referential to the word *vopros* 'question'.

5 Corpus annotation

The first text annotated for SemOntoCor is the story "The Little Prince" by Antoine de Saint-Exupery, translated into Russian by Nora Gal. It was first published in French in 1943 and translated into more than 180 languages. It is one of the most popular books in the world literature, with over 80 million copies sold. The (Russian translation of the) text contains 1532 sentences (13120 tokens). It is included in at least two other known semantic corpora – AMR (https://amr.isi.edu/download.html) and UNL [Martins 2012]. It offers a rare opportunity to compare and assess different variants of semantic markup, as well as to develop a procedure of automatic (or semi-automatic) translation from one mark-up to another. The latter is interesting in that it opens up an enticing prospect of supplementing SemOntoCor with the data from other corpora.

As of today (May 2023), more than 1100 sentences of the story have been marked up. When the story is fully annotated, it will be made available to the public.

5.1 The BSemS format

The format in which BSemSs are stored is an extension of the XML format used for representing SynTagRus – the treebank of Russian which is an integral part of the National Corpus of Russian (https://ruscorpora.ru). The format used for SynTagRus is described in [Iomdin, Sizov 2009] and its extension for SemOntoCor is proposed in [Frolova, Rygaev 2022].

Each sentence and all linguistic information about it are stored inside the <S> tag, which contains <W> tags for the words of the sentence. They are followed by the <SEM> tag for the semantic structure.

Each <W> contains the information about the ID number of the word, its morphological features and incoming syntactic links.

In <SEM> tag the information is organized into <N> tags, each referring to a certain semantic node.

Each <N> tag has several attributes, ID, TYPE, and VALUE. ID attribute gives the number of the node, TYPE attribute shows the type of the node with respect to ontology, see more in [Frolova, Rygaev 2022]. VALUE attribute contains the name of the node.

If a node has outgoing links, they are listed in the <R> tag, which has attributes LINK (for the name of the link) and TO (for the ID of the target node).

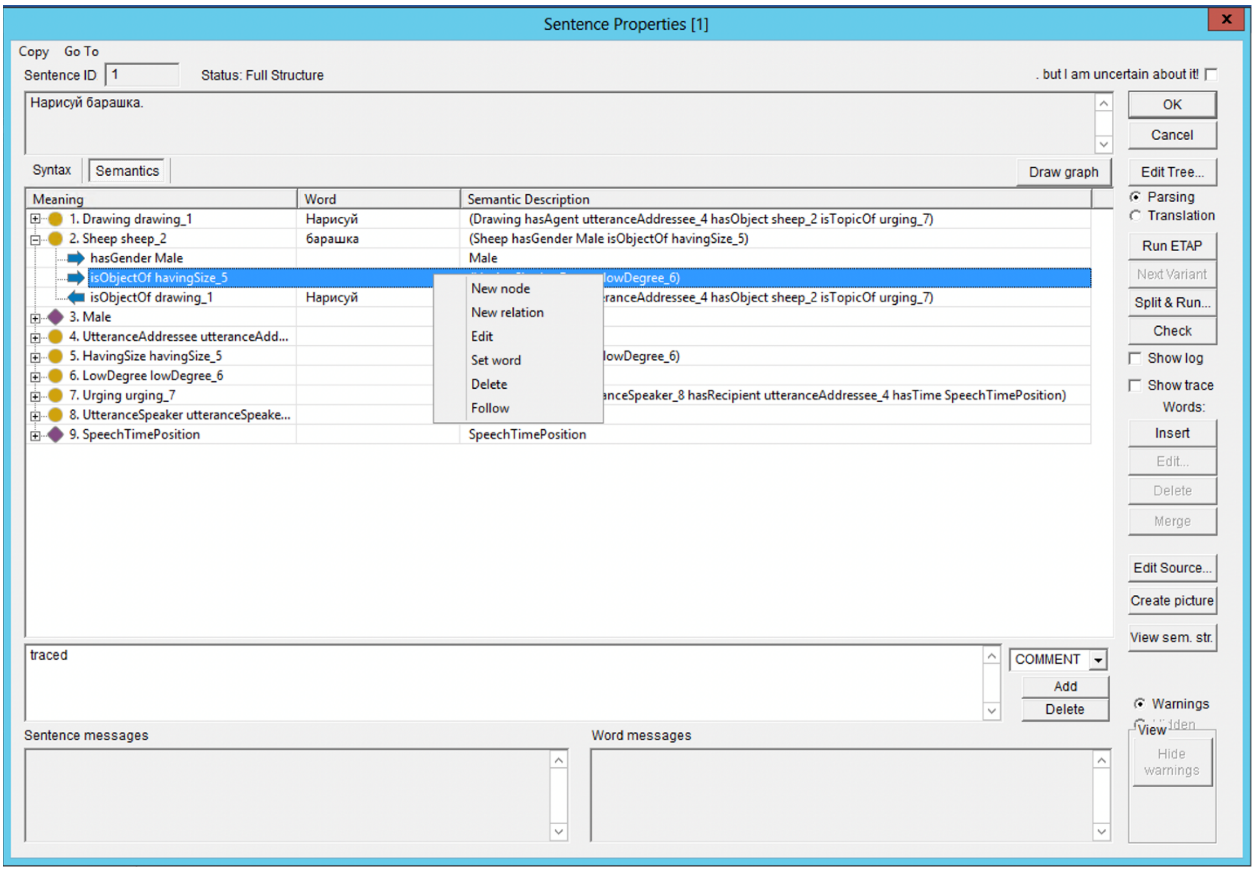
Below is the BSemS of the sentence *Narisuj barashka* 'Draw a lamb'. It has two <W> tags according to the number of words in the sentence and nine <N> tags inside <SEM> according to the number of nodes in BSemS. For example, the node with ID="2" has value "Sheep", is connected to nodes 3 and 5 by links hasGender and isObjectOf respectively.

```
<S COMMENT="traced" DATE="09 02 2023 14:38:09" ID="1"
        <W DOM="_root" EXTRAFEAT="САР ЛИЧ" FEAT="V СОВ ПОВ ЕД 2-Л" ID="1"
KSNAME="РИСОВАТЬ" LEMMA="РИСОВАТЬ">Нарисуй</W>
        <W DOM="1" FEAT="S ЕД МУЖ ВИН ОД" HYPOT="1-компл.11" ID="2"
KSNAME="БАРАШЕК1" LEMMA="БАРАШЕК" LINK="1-компл">барашка</W>.
        <SEM>
                <N ID="1" TYPE="anonymous" VALUE="Drawing">
                        <R LINK="hasAgent" TO="4"/>
                        <R LINK="hasObject" TO="2"/>
                        <R LINK="isTopicOf" TO="7"/>
                </N>
                <N ID="2" TYPE="anonymous" VALUE="Sheep">
                        <R LINK="hasGender" TO="3"/>
                        <R LINK="isObjectOf" TO="5"/>
                </N>
                <N ID="3" TYPE="named" VALUE="Male"/>
                <N ID="4" TYPE="anonymous" VALUE="UtteranceAddressee"/>
                <N ID="5" TYPE="anonymous" VALUE="HavingSize">
                        <R LINK="hasDegree" TO="6"/>
                </N>
                <N ID="6" TYPE="anonymous" VALUE="LowDegree"/>
                <N ID="7" TYPE="anonymous" VALUE="Urging">
                        <R LINK="hasAgent" TO="8"/>
                        <R LINK="hasRecipient" TO="4"/>
                        <R LINK="hasTime" TO="9"/>
                </N>
                <N ID="8" TYPE="anonymous" VALUE="UtteranceSpeaker"/>
                <N ID="9" TYPE="named" VALUE="SpeechTimePosition"/>
        </SEM>
</S>
```

5.2 Annotation tool

To annotate SemOntoCor we use a custom Structure Editor (StrEd) software, originally developed for annotation of the SynTagRus treebank [Iomdin,Sizov, 2009] and later extended in order to annotate SemOntoCor. StrEd supports two annotation procedures – manual and semi-automatic. Fig. 1 shows sentence *Narisuj barashka* loaded into StrEd and its BSemS obtained automatically.

In the manual mode, the annotator enters each node individually and connects them by semantic relations using the context menu, shown in Fig. 1. The menu contains options of creating and deleting nodes and links between them, as well as changing the correspondence between nodes and words. In the semi-automatic procedure, the annotator runs the SemETAP semantic parser for each sentence (cf. section 3 above). The BSemS obtained is loaded into StrEd for subsequent revision.



Sentence ID 1 Status: Full Structure . but I am uncertain about it! □

Нарисуй барашка.

Meaning	Word	Semantic Description
1. Drawing drawing_1	Нарисуй	(Drawing hasAgent utteranceAddressee_4 hasObject sheep_2 isTopicOf urging_7)
2. Sheep sheep_2	барашка	(Sheep hasGender Male isObjectOf havingSize_5)
hasGender Male		Male
isObjectOf havingSize_5		lowDegree_6)
isObjectOf drawing_1	Нарисуй	ranceAddressee_4 hasObject sheep_2 isTopicOf urging_7)
3. Male		
4. UtteranceAddressee utteranceAdd...		
5. HavingSize havingSize_5		lowDegree_6)
6. LowDegree lowDegree_6		
7. Urging urging_7		nceSpeaker_8 hasRecipient utteranceAddressee_4 hasTime SpeechTimePosition)
8. UtteranceSpeaker utteranceSpeake...		
9. SpeechTimePosition		SpeechTimePosition

● Parsing ○ Translation

□ Show log

□ Show trace

traced

COMMENT

● Warnings ○ View hidden

Figure 1: StrEd annotation tool

In the sentence properties window we can toggle between “Syntax” and “Semantics” tabs (upper left), which allows the user to choose between syntactic and semantic annotation.

The Semantics tab opens the table, where each line corresponds to a BSemS node. For each node the following information is available: the node number and its name, the word of the sentence which generated this node (if any) and a semantic description of the node, which is a fragment of BSemS containing outgoing links and nearest dependents.

A node can be expanded if it is connected to other nodes either with an incoming or an outgoing link. Node 2 (Sheep) in Fig.1 is expanded to reveal two outgoing links (blue arrows pointing to the right) and an incoming link (arrow pointing to the left) from Drawing.

For the annotator’s convenience, the “Draw graph” button in the upper right corner visualizes the BSemS. Fig. 2 shows the visualization of BSemS in Fig. 1.

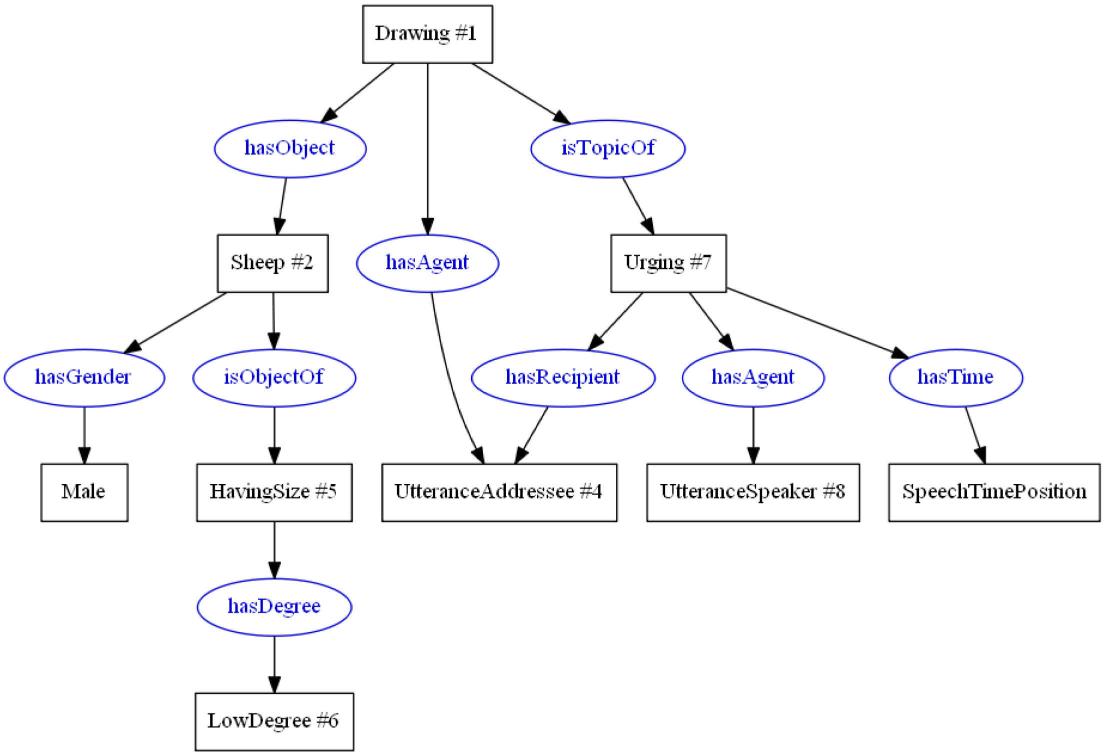


Figure 2: Visualization of the BSemS for *Narisuj barashka*

We performed a rough estimation of the current speed of annotation by selecting arbitrarily 10 sentences of average length (9-17 tokens) and complexity. The same annotator annotated 5 of them manually and the other 5 semi-automatically. The annotation proper (i.e. without the time spent on thinking) took 11,6 minutes/sentence for manual annotation and 8,4 minutes/sentence for semi-automatic annotation. New annotators, that will be recruited for the job, will undergo special training. At first, their annotation speed will probably be lower. However, given that annotators will acquire experience over time and the SemETAP parser will gain accuracy, we can expect the speed of annotation to increase both in manual and semi-automatic mode.

6 Conclusion

We present SemOntoCor - a new semantic corpus for Russian under construction at the Institute for Information Transmission Problems (RAS). It is the next step in the development of SynTagRus with its several types of annotation. SemOntoCor builds on top of the morpho-syntactic annotation of SynTagRus and assigns each sentence a Basic Semantic Structure (BSemS). BSemS represents the direct meaning of the sentence in terms of ontological concepts and semantic relations between them. It abstracts away from lexico-syntactic variation and in many cases decomposes lexical meanings into smaller elements. SemOntoCor is unified with the SemETAP semantic parser, which produces two levels of semantic structures – Basic SemS and Enhanced SemS. The latter enriches BSemS with different types of inferences, based both on the linguistic and the common-sense knowledge. The annotation is done in a semi-automatic mode: SemETAP produces a draft BSemS, which is then revised by an expert. In the first version of SemOntoCor a novel by Antoine de Saint-Exupery “The Little Prince” is annotated.

Acknowledgements

This work was done with the financial support of the grant No. 22-28-01941 “Development of the infrastructure and the first phase of the semantic corpus for Russian” from the Russian Science Foundation.

References

[1] 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.

[2] Abzianidze, L., J. Bjerva, K. Evang, H. Haagsma, R. van Noord, P. Ludmann, D.-D. Nguyen, and J. Bos (2017, April). The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, pp. 242–247. Association for Computational Linguistics.

[3] Abzianidze L. and Johan Bos. (2017). Towards universal semantic tagging. In: Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers, pages 1–6, Montpellier, France.

[4] Apresian Ju., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L. (2003). ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // First International Conference on Meaning-Text Theory (MTT’2003). June 16-18, 2003. Paris: Ecole Normale Supérieure. P. 279-288.

[5] Bada M., Eckert M., Evans D., Garcia K., Shipley K., Sitnikov D., Baumgartner W.A. Jr, Cohen K.B., Verspoor K., Blake J.A., Hunter L.E. (2012) Concept annotation in the CRAFT corpus. // BMC Bioinformatics. 2012 Jul 9 13:161. DOI: 10.1186/1471-2105-13-161. PMID: 22776079; PMCID: PMC3476437.

[6] Banarescu L., Bonial C., Shu Cai, Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn Ph., Palmer M., Schneider N. (2013) Abstract meaning representation for sembanking. // Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Sofia, Bulgaria. pp. 178–186.

[7] Baker C.F., Fillmore C., Lowe J.B. (1998) The Berkeley FrameNet project. // Proceedings of COLING-ACL, Montreal, Canada.

[8] Basile, V., J. Bos, K. Evang, and N. Venhuizen (2012). Developing a large semantically annotated corpus. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp. 3196 – 3200.

[9] Bauer D., Fürstenau H., Rambow O. (2012) The Dependency-Parsed FrameNet Corpus. // In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp 3861–3867. European Language Resources Association (ELRA).

[10] Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. (2016) Annotating the Little Prince with Chinese AMRs. // In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pages 7–15, Berlin, Germany. Association for Computational Linguistics.

[11] Bjerva J., Barbara Plank, and Johan Bos. (2016). Semantic tagging with deep residual networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3531–3541, Osaka, Japan.

[12] Boguslavsky I. (2017). Semantic Descriptions for a Text Understanding System. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017” Moscow, May 31—June 3, 201

[13] Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S. (2018). Semantic analysis with inference: high spots of the football match. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”, Moscow, May 30—June 2, 2018

[14] Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S. (2019). Knowledge-based approach to Winograd Schema Challenge. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. Moscow, May 29—June 1, 2019.

[15] Boguslavsky I., Dikonov V., Inshakova E., Iomdin L., Lazursky A., Rygaev I., Timoshenko S., Frolova T. (2021) Semantic Representations in Computational and Theoretical Linguistics: the Potential for Mutual Enrichment. pp. 127-141. DOI 10.28995/2075-7182-2021-20-127-141.

[16] Bos J., Abzianidze L. (2019). Thirty Musts for Meaning Banking. In Proceedings of the First International Workshop on Designing Meaning Representations. — Florence, Italy, August 1st. — 2019. — p. 15–27.

[17] Bos J., Basile V., Evang K., Venhuizen N.J., Bjerva J. (2017) The Groningen Meaning Bank. // In book: Handbook of Linguistic Annotation, pp.463-496, Springer Netherlands DOI:10.1007/978-94-024-0881-2_18

[18] Burchardt A., Pennacchiotti M. (2008) FATE: a FrameNet-Annotated Corpus for Textual Entailment. // In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

[19] Carlson L., Marcu D., Okurowski M. E. (2001). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001. Aalborg, Denmark, 1–10.

[20] Fellbaum C., Landes S., Leacock C. (1998) Building semantic concordances. // In Fellbaum (1998), chapter 8, pp 199–216.

[21] Fillmore C. J., (1976) Frame semantics and the nature of language. // Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280:20–32.

[22] Flickinger D., Yi Zhang, and Valia Kordoni. 2012. Deepbank: A dynamically annotated treebank of the Wall Street journal. In Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories, pages 85–96. Edicoes Colibri.

[23] Frolova T.I., Rygaev I.P. (2022), Razrabotka formata razmetki i printsipov annotirovanija dlja semanticheskogo korpusa russkogo jazyka na materiale “Malen’kogo printsa” [Developing a mark-up format and annotation principles for a semantic corpus of Russian on the material of The Little Prince]. In: Sbornik trudov 46 mezhdistsiplinarnoj shkoly-konferentsii IPPI RAN “Informatsionnye texnologii i sistemy 2022 (ITiS'2022)”. M. IPPI, p. 122-136.

[24] Hajič J. (2002) 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. pp. 216–226.

[25] Hajič J., Hladká B., Pajas P. (2001) The Prague Dependency Treebank: Annotation Structure and Support. // IRCS Workshop on Linguistic Databases. pp. 105–114.

[26] Hovy E., Mitchell M., Palmer M., Ramshaw L., Weischedel R. (2006). *OntoNotes*: The 90% Solution. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 57–60, New York, June 2006.

[27] Inshakova E.S., Iomdin L.L., Mitjushin L.G., Sizov V.G., Frolova T.I., Cinman L.L. (2019), SinTagRus segodnja [The SynTagRus Today]. In: Trudy Instituta russkogo jazyka im. V.V. Vinogradova [Proceedings of Vinogradov Russian Language Institute]. Vol. 21, Moscow, pp. 14–41. DOI: 10.31912/pvrli-2019.21.1.

[28] Iomdin L., Sizov V. Structure Editor: a Powerful Environment for Tagged Corpora // MONDILEX Fifth Open Workshop. Ljubljana, Slovenia, October 14–15, 2009. Ljubljana, 2009. P. 1-12. ISBN 978-961-264-012-5.

[29] Kamp H., Reyle U. (1993) From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT. // Kluwer, Dordrecht.

[30] Kipper K., Korhonen A., Ryant N., Palmer M. (2008) A large-scale classification of English verbs. // Language Resources and Evaluation Journal, 42:21–40

[31] Kustova G., Ljashevskaja O., Paducheva E., Raxilina E. (2005). Semanticheskaja razmetka leksiki v Natsional’nom korpuse russkogo jazayka: printsipy, problemy, perspektivy. [Semantic mark-up of words in Russian National Corpus: principles, problems, perspectives]. In: Natsional’nyj korpus russkogo jazyka: 2003-2005. Rezultaty i perspektivy. M., pp. 155-174.

[32] Martins R.T. (2012) Le Petit Prince in UNL. // International Conference on Language Resources and Evaluation (2012).

[33] Mann W.C., Thompson S.A. (1988) Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. // Text 8(3), pp 243–281.

[34] Martins, R. (2012). Le Petit Prince in UNL. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 3201–3204, Istanbul, Turkey. European Language Resources Association (ELRA).

[35] McShane M., S. Nirenburg. (2021). Linguistics for the Age of AI. The MIT Press. Cambridge, Massachusetts.

[36] 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.

[37] Mel’čuk I. Semantics: From Meaning to Text. Vol. 1. — Amsterdam/Philadelphia: John Benjamins. — 2012.

[38] Mel’čuk I. Semantics: From Meaning to Text. Vol. 2. — Amsterdam/Philadelphia: John Benjamins. — 2013.

[39] Mel’čuk I. Semantics: From Meaning to Text. Vol. 3. — Amsterdam/Philadelphia: John Benjamins. — 2015.

[40] Meyers A., Reeves R., Macleod C., Szekely R., Zielinska V., Young B., Grishman R., (2004) The NomBank Project: An Interim Report. // In proceedings of FCP@NAACL-HLT (2004).

[41] Oepen, S., D. Flickinger, K. Toutanova, and C. D. Manning (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. Research on Language and Computation 2(4), 575 – 596.

[42] O'Gorman, T. J., Regan M., Griffitt K., Hermjakob U., Knight K., Palmer M. (2018) AMR Beyond the Sentence: the Multi-sentence AMR corpus. // International Conference on Computational Linguistics (2018).

[43] Palmer M. (2002) From Treebank to PropBank. // In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain.

[44] Palmer M., Kingsbury P., Gildea D. (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. // Computational Linguistics. 31 (1): 71–106. CiteSeerX 10.1.1.136.8985. doi:10.1162/0891201053630264. S2CID 2486369.

[45] Ruppenhofer J., Ellsworth M., Petruck M.R.L, Johnson C.R, Baker C.F., Scheffczyk J. (2007) FrameNet II: Extended Theory and Practice // Book (Revised November 1, 2016). Available at https://framenet.icsi.berkeley.edu/fndrupal/the_book

[46] Petrolito T., Bond F. (2014) A Survey of WordNet Annotated Corpora. // Global WordNet Conference 2014.

[47] Raxilina E., Kustova G., Ljashevskaja O., Reznikova T., Shemanaeva O. (2009). Zadachi i printsipy semanticheskoj razmetki leksiki v NRRJa [Tasks and principles of the semantic mark-up of words in RNC]. In: Natsional'nyj korpus russkogo jazyka: 2006-2008. Rezultaty i perspektivy. Spb.: Nestor-Istorija, pp. 215-239.

[48] Taboada M., Mann W.C. (2006) Rhetorical Structure Theory: Looking Back and Moving Ahead. // Discourse Studies 8, pp. 423–459.

[49] Timoshenko S.P., Iomdin L.L., Gladilin S.A., Inshakova E.S. (2021), SinTagRus v sostave NKRJa: novye vozmozhnosti [The SynTagRus as part of RNC: new opportunities]. In: Trudy mezhdunarodnoj konferentsii "Korpusnaja lingvistika-2021" [Proceedings of the International conference "Corpus linguistics-2021"], p.31-43.

[50] Weischedel R., Hovy E., Mitchell M., Palmer M., Belvin R., Pradhan S., Ramshaw L., Xue N. (2009). OntoNotes: A Large Training Corpus for Enhanced Processing // In J. Olive, C. Chrisiansen, and J. McCrary (eds), Handbook of Natural Language Processing and Machine Translation.

[51] White A.S., Stengel-Eskin E., Vashishtha S., Govindarajan V.S., Reisinger D.A., Vieira T., Sakaguchi K., Zhang S., Ferraro F., Rudinger R., Rawlins K., Van Durme B.. (2020) The Universal Decompositional Semantics Dataset and Decomp Toolkit. // In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 5698–5707, Marseille, France. European Language Resources Association.

[52] Zeldes A. (2017) The GUM Corpus: Creating Multilayer Resources in the Classroom. // Language Resources and Evaluation 51(3), 581–612.

[53] Zeman D., Hajic J. (2020) FGD at MRP 2020: Prague Tectogrammatical Graphs. // In Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, pp. 33–39, Online. Association for Computational Linguistics.

[54] Zhao M., Wang Y., Lepage Y., (2022) Large-scale AMR Corpus with Re-generated Sentences: Domain Adaptive Pre-training on ACL Anthology Corpus // International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2022, pp. 19-24, DOI: 10.1109/ICACSIS56558.2022.9923502.