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

Строение и особенности разметки корпуса семантических структур SemOntoCor[1]

Аннотация. Данная работа представляет корпус семантических структур для русского языка, создаваемый в Лаборатории компьютерной лингвистики ИППИ РАН. Корпус на сегодняшний день включает около 5000 предложений. Семантическая разметка фраз соотносится с разметкой текстов на других уровнях: морфологическом, синтаксическом, лексико-функциональном, микросинтаксическом.

Ключевые слова. семантический корпус; глубоко аннотированный корпус; онтология, семантика

1. Общие сведения о корпусе

В Лаборатории компьютерной лингвистики ИППИ РАН с 2022 г. ведется работа по составлению семантического корпуса русского языка SemOntoCor (= Семантический Корпус, опирающийся на Онтологию). Он представляет собой набор текстов, в котором каждому предложению сопоставлена его семантическая структура. SemOntoCor составляется не с нуля, а представляет собой следующий шаг в развитии многослойного синтаксического корпуса СинТагРус, входящего в состав НКРЯ (https://ruscorpora.ru/search/syntax). Поэтому SemOntoCor содержит также всю аннотацию, имеющуюся в СинТагРусе (морфологическую, синтаксическую, лексическую, лексико-

1 Работа выполнена в рамках государственного задания МинОбрНауки

функциональную, микросинтаксическую, кореферентную, метатекстовую). Подробнее о СинТагРусе и видах разметки в нем см. в [Богуславский и др. 2024]. Благодаря этому появляется возможность на одном и том же текстовом материале и на едином теоретическом базисе получить ресурс, имеющий несколько слоев разметки, начиная от морфологической структуры и заканчивая семантической.

Корпус создается в рамках семантического проекта [Богуславский 2021], теоретической основой которого являются модель «Смысл ⇔ Текст» И.А. Мельчука [Мельчук 1999] и теория интегрального описания языка Ю.Д. Апресяна [Апресян 1995].

1.1. Объем и состав корпуса

Объем корпуса SemOntoCor на апрель 2025 г. – около 5000 предложений. Сейчас в SemOntoCor входят следующие тексты:

- сказка «Маленький принц» А. де Сент-Экзюпери в переводе на русский Н. Галь.
- две статьи из русской Википедии – «Альпинизм» (версия 2011 г.) и «Поведение» (версия 2013 г.). Общий объем двух текстов – 545 предложений.
- публицистические и научно-популярные журнальные статьи (более 2000 предложений).
- материалы новостных лент 2000-2002 гг. (467 предложений).

2. Разметка предложения

Структуры имеют несколько форматов представления: в виде троек RDF, на формальном языке Эталог, специально разработанном для записи лингвистами семантических структур и правил [Rygaev 2018], и в виде наглядной картинки.

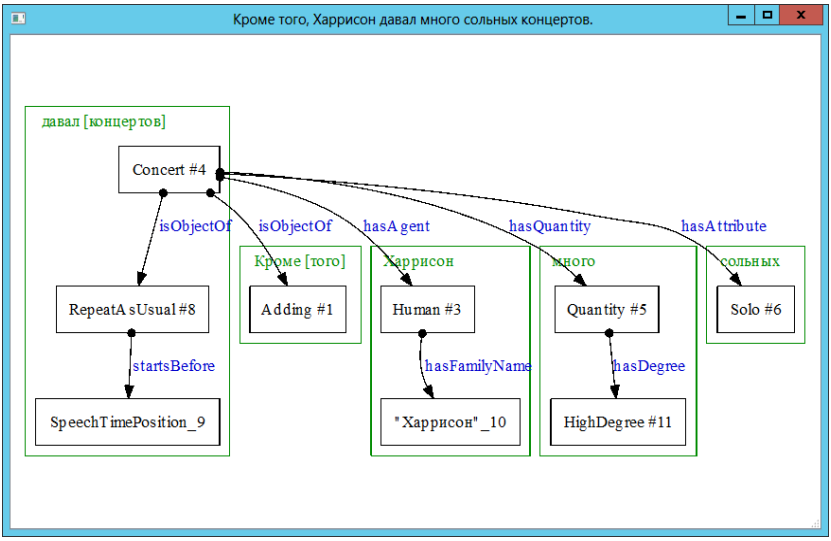


Рис. 1. **Семантическая структура фразы *Кроме того, Харрисон давал много сольных концертов.***

Ниже приведено семантическое представление одного из предложений корпуса на языке Эталог, а выше на рис. 1 – визуализация этого представления.

Кроме того, Харрисон давал много сольных концертов.

Concert concert_4 // *давал [концертов]*
    isObjectOf (RepeatAsUsual repeatAsUsual_8
        startsBefore SpeechTimePosition_9)
    hasAgent (Human human_3 // *Харрисон*
        hasFamilyName “Харрисон”_10)
    hasQuantity (Quantity quantity_5 // *много*
        hasDegree HighDegree highDegree_21)
    hasAttribute Solo solo_6 // *сольных*
    isObjectOf Adding adding_1 // *Кроме [того]*

Структура читается примерно так: «Имели место концерты, повторявшиеся в прошлом, активным участником которых был человек по фамилии Харрисон, количество которых было большим, которые были сольными, утверждение о концертах является объектом добавления (говорящим)».

2.1. Строение семантической структуры

Семантическая структура каждого предложения представляет собой граф, в узлах которого стоят семантические индивиды и литералы (строки и цифры), а ребрами являются семантические отношения. Отношения и индивиды хранятся в онтологии OntoETAP. Эта онтология построена на базе SUMO (Suggested Upper Merged Ontology; ontologyportal.org), которая была значительно переработана с учетом потребностей адекватного лингвистического описания. В ней содержится концептуализированная информация о мире. Два типа онтологических единиц, используемых при построении семантических структур – это отношения и индивиды.

При построении базовых семантических структур используется около сотни отношений, их имена легко интерпретируются в рамках традиционной семантики: hasAgent, hasObject, hasTime и т.д. Актуальный список отношений доступен по адресу https://docs.google.com/document/d/1XxhnnhRqUpTvPZabi4qesgBoCRAwzCitg_KfF4ABIr4/edit?tab=t.0.

В узлах семантического графа стоят чаще всего индивиды (инстанции) онтологических классов. Классов в онтологии несколько тысяч, список их открыт. Классы образуют иерархию, при этом свойства надклассов наследуются подклассами. Наследуемые свойства формулируются также в терминах онтологических отношений и классов. Так, онтологический класс Tying,

онтокоррелят русского слова ПРИВЯЗЫВАТЬ, имеет собственное свойство hasInstrument String (иметь инструментом объект класса String – класса длинных тонких объектов: веревка, нить, волос и т.д.). Кроме собственных свойств Tying наследует свойство своего надкласса Attaching: hasAgent Agent (иметь активного деятеля – лицо), а также hasObject Object и hasObject2 Object (иметь первый и второй объекты, на которые направлено действие).

3. Некоторые принципы разметки

При построении базовых семантических структур соблюдается несколько принципов.

Один из таких принципов состоит в **учете направления синтаксических связей** в анализируемом предложении. В большинстве случаев направление семантических связей между узлами в базовых структурах совпадает с направлением синтаксических связей между соответствующими словами. Для того, чтобы обеспечить такое совпадение, используется механизм инверсных отношений. Большинство семантических отношений из списка, ссылка на который приведена выше в разделе 2.1, имеют инверсные соответствия: для hasAgent инверсным отношением является isAgentOf, для hasTime – isTimeOf и т.д. Семантически записи Meeting hasTime Morning и Morning isTimeOf Meeting эквивалентны, но первую запись удобно использовать для описания смысла фразы *утренняя встреча*, а вторую – для фразы *утро встречи*.

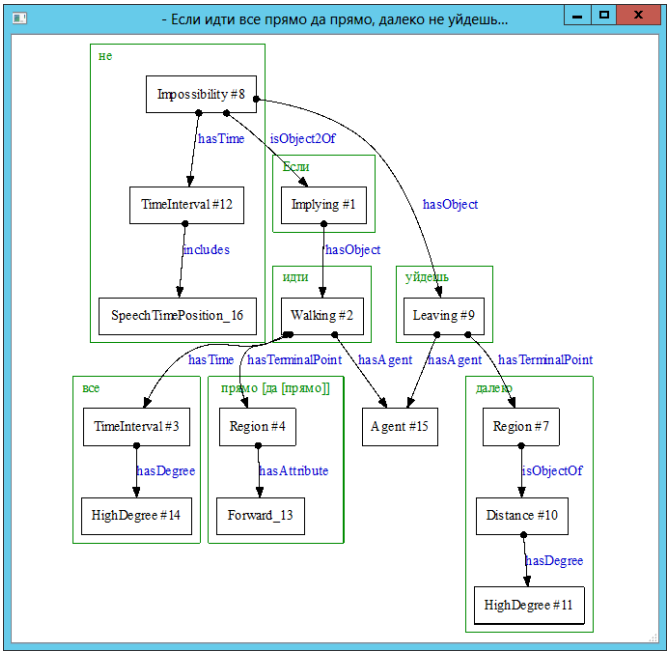


Рис. 2. **Семантическая структура фразы *Если идти все прямо да прямо, далеко не уйдешь.***

Другой принцип состоит том, что в семантической структуре обязательно должны присутствовать **сведения о заполнении валентностей** семантических единиц, если эти сведения имеются в предложении. В предложении, проиллюстрированном на рис. 2, валентности активного участника у узлов Walking и Leaving (соответствующих русским словам *идти* и *уйдешь*) не заполнены в предложении, но тем не менее присутствует знание от кореферентности этих активных участников. Для того, чтобы отразить это, в базовую структуру вводится узел, обозначенный на рисунке как Agent#15, который не соответствует никакому слову в предложении, но

обеспечивает учет кореферентности активных участников обеих ситуаций.

4. Техника разметки

Разметка семантического корпуса может вестись в двух режимах – полностью **вручную** и в **полуавтоматическом** режиме. В **ручном** режиме аннотатор строит с нуля семантическую структуру на языке Эталог (об этом формальном языке см. [Rygaev 2018]) и помещает ее в редактор структур в специальное окно, предназначенное для работы с этим языком. При разметке в **полуавтоматическом** режиме используется структура, полученная в результате работы правил перевода. Эту структуру можно редактировать в **табличном** режиме, а также **на языке Эталог** или **в графическом редакторе**.

Табличный режим на данный момент почти не используется, т.к. теперь в двух других режимах редактировать семантическую разметку стало проще и быстрее.

Редактирование на языке Эталог происходит в том же окне редактора структур, которое используется для внесения ручной разметки. В режиме редактирования при открытии окна текст на Эталоге автоматически формируется из внутреннего представления семантической структуры. Его можно отредактировать вручную. После сохранения он преобразуется обратно во внутреннее представление. Для задания связи семантических узлов со словами размечаемого предложения установлено соглашение, что номер узла должен соответствовать номеру слова в предложении. Для большего удобства слева отображается список слов предложения с их номерами, а при автоматическом формировании текста на Эталоге узлы

снабжаются комментариями, в которых указаны слова или фразы, которым они соответствуют (см. рис. 3).

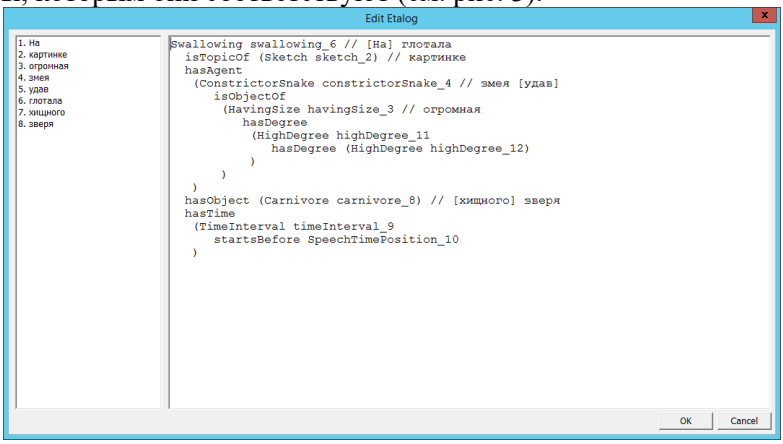


Рис. 3. **Окно редактирования на Эталоге со структурой фразы *На картинке огромная змея – удав – глотала хищного зверя.***

В графическом режиме семантическая структура изображается в виде интерактивной картинки графа, где можно выделять отдельные элементы (узлы, стрелки) и выполнять с ними различные действия. Например, у узла можно поменять семантический концепт и его привязку к слову предложения, а у стрелки – выбрать входящий и исходящий узел и исправить метку – наименование отношения из онтологии.

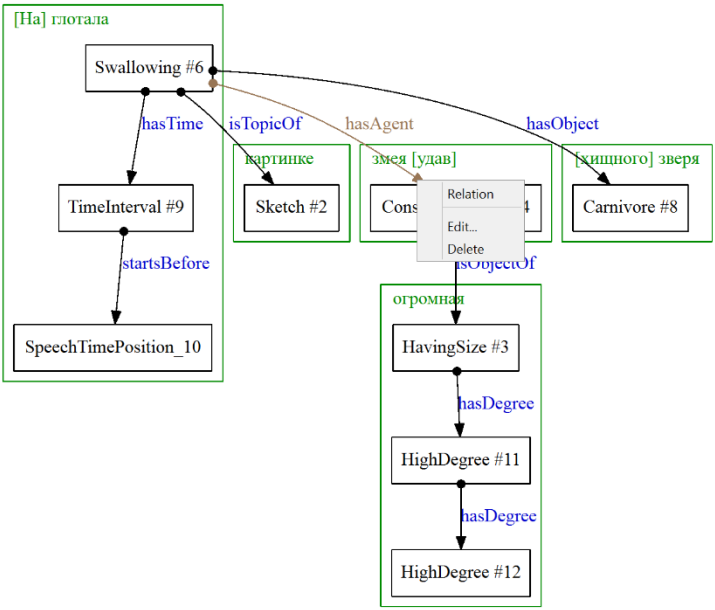


Рис. 4. **Окно графического режима редактирования со структурой фразы *На картинке огромная змея – удав – глотала хищного зверя.***

В графическом редакторе есть возможность добавления новых объектов, отношений и узлов (см. рис. 4), доступно масштабирование изображения.

Одновременно с разметкой происходит правка и пополнение онтологии. При работе в полуавтоматическом режиме происходит также отладка правил конвертации синтаксических структур в семантические.

5. Заключение

В статье были рассмотрены некоторые принципы и приемы построения семантического корпуса SemOntoCor. Семантическая разметка создает материал для исследования

соотношения смыслов и синтаксических (и морфологических) способов их выражения.

Литература

1. Апресян Ю.Д. (1995), Интегральное описание языка и системная лексикография. М.: Школа «Языки русской культуры».

2. Богуславский И.М. (2021), Семантический анализ с опорой на умозаключения в функциональной модели языка. Вопросы Языкознания, 2021. №1. С. 29-56.

3. Богуславский И.М., Дяченко П.В., Иншакова Е.С., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Мовсесян А.А., Рыгаев И.П., Сизов В.Г., Тимошенко С.П., Фролова Т.И., Чага А.В. (2024), Современное состояние корпуса СинТагРус. Труды Института Русского языка им. В.В. Виноградова. № 4 (42), с. 141-169.

4. Мельчук И.А. (1999), Опыт теории лингвистических моделей «Смысл ⇔ Текст». М.: Школа «Языки русской культуры».

5. Rygaev I.P. (2018), Etalog – a natural-looking knowledge representation formalism. Труды школы-конференции ИТиС 2018, с. 473-482.

References

1. Apresjan Ju.D. (1995), Integral’noje opisanije jazyka i sistemnaja leksikografija.. Moscow: Shkola “Yazyki russkoy kultury”.

2. Boguslavsky I.M. (2021), Семантический анализ с опорой на умозаключения в функциональной модели языка. Вопросы Языкознания, 2021. №1. С. 29-56.

3. Boguslavsky I.M., Djachenko P.V., Inshakova E.S., Iomdin L.L., Lazursky A.V., Mit’ushin L.G., Movsesjan A.A., Rygaev I.P., Sizov V.G., Timoshenko S.P., Frolova T.I., Chaga A.V.

(2024), Sovremennoje sostojanije korpusa SynTagRus. Trudy Instituta Russkogo jazyka im. V.V. Vinogradova № 4 (42), p. 141-169.

4. Mel'chuk I.A. (1999), Opyt teorii lingvisticheskih modelej «Smysl–Tekst». Moscow: Shkola "Yazyki russkoy kultury".

5. Rygaev I.P. (2018), Etalog – a natural-looking knowledge representation formalism. Trudy shkoly-konferentsii ITaS 2018, p. 473-482.

Structure and Tagging of the SemOntoCor Semantic Corpus

I.M. Boguslavskij, V.G. Dikonov, E.S. Inshakova,
A.V. Lazurskij, I.P. Rygaev, S.P. Timoshenko, T.I. Frolova

Abstract. The paper presents the project of a semantic corpus named SemOntoCor. For April 2025, SemOntoCor contains 5000 sentences from Russian texts tagged with semantic graphs that represent the meaning of these sentences. All texts of SemOntoCor are also present in SynTagRus, thus the semantic structure of each sentence is always accompanied by its morphological and syntactic structure and information about Lexical Functions and Microsyntax within it. The paper deals with some principles of semantic tagging and the ways and instruments used for implementation of these principles. The OntoETAP Ontology is mentioned as a source of elements of which the semantic graphs are constructed. Some examples of the structures are presented as visualized graphs.

Keywords. semantic corpus; deeply annotated corpus; ontology, semantics

---

Богуславский Игорь Михайлович
Институт проблем передачи информации РАН (Россия)
Мадридский политехнический университет (Испания)
Boguslavsky Igor

Institute for Information Transmission Problems RAS (Russia)
Universidad Politécnica de Madrid (Spain)
E-mail: bogus@iitp.ru

Диконов Вячеслав Григорьевич
Институт проблем передачи информации РАН, (Россия)
Vjacheslav G. Dikonov
IITP Kharkevich Institute (Russia)
E-mail: sdiconov@mail.ru

Иншакова Евгения Сергеевная
Институт проблем передачи информации РАН, (Россия)
НИУ "Высшая школа экономики"
Inshakova Evgeniya
IITP Kharkevich Institute (Russia)
HSE University
E-mail: e.s.inshakova@gmail.com

Лазурский Александр Вадимович
Институт проблем передачи информации РАН, (Россия)
Alexandre V. Lazursky
IITP Kharkevich Institute (Russia)
E-mail: lazursky@mail.ru

Рыгаев Иван Петрович
Университет Тюбингена (Германия)
Ivan Rygaev
University of Tübingen (Germany)
E-mail: irygaev@gmail.com

Тимошенко Светлана Петрровна
Институт проблем передачи информации РАН, (Россия)
Svetlana P. Timoshenko
IITP Kharkevich Institute (Russia)
E-mail: nyrestein@gmail.com

Фролова Татьяна Ильинична
Институт проблем передачи информации РАН, (Россия)
Tatiana I. Frolova
IITP Kharkevich Institute (Russia)
E-mail: tfrolova@gmail.com