Ю.Д. Апресян, Л.Л. Иомдин, А.В. Санников, В.Г. Сизов

СЕМАНТИЧЕСКАЯ РАЗМЕТКА В ГЛУБОКО АННОТИРОВАННОМ КОРПУСЕ РУССКОГО ЯЗЫКА

В данной работе речь идет об обогащении семантической информацией аннотированного корпуса русских текстов, который разрабатывается Лабораторией компьютерной лингвистики ИППИ РАН (руководитель – И.М. Богуславский) в сотрудничестве с Сектором теоретической семантики ИРЯ РАН (руководитель – Ю.Д. Апресян).

Работа состоит из двух частей. В первой обсуждается язык для семантической разметки, а во второй – технические вопросы введения семантической информации в реально существующий синтаксически аннотированный корпус русских текстов.

Этот корпус основан на идеологии системы машинного перевода ЭТАП[1]. В настоящий момент он является частью исследовательского проекта общероссийского масштаба – Национального корпуса русского языка (**www.ruscorpora.ru**) – и единственным русскоязычным корпусом, содержащим синтаксическую разметку. Ценность корпуса определяется прежде всего глубиной разметки: каждому слову приписывается исчерпывающая морфологическая информация, а для каждого предложения строится полное синтаксическое дерево зависимостей. Приведем пример такого размеченного предложения (см. рис. 1).

---

[1] *Boguslavsky I.M., Grigorieva S.A, Grigoriev N.V, Kreidlin L.G., Frid N.E.* Dependency Treebank for Russian: Concepts, Tools, Types of Information // Proceedings of the 18th Conference on Computational Linguistics. Vol. 2. Saarbrücken, 2000. P. 987–991; *Богуславский И.М., Григорьев Н.В., Григорьева С.А., Иомдин Л.Л., Крейдлин Л.Г., Фрид Н.Е.* Разработка синтаксически размеченного корпуса русского языка // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб., 2002. С. 40–50.

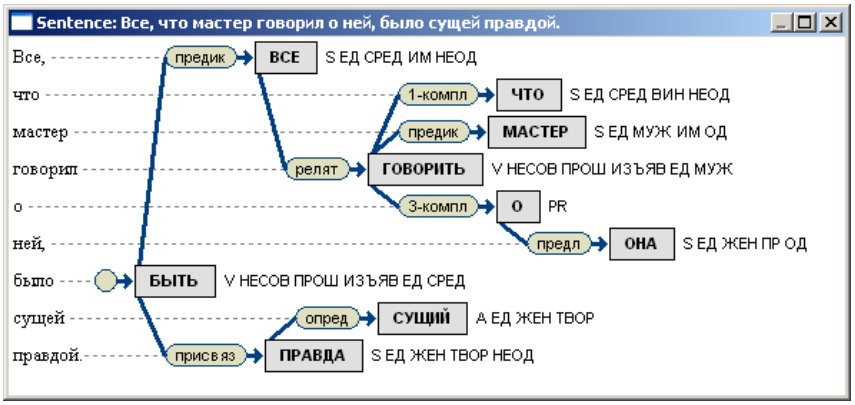

Рис. 1. Пример размеченного предложения

Как видно из примера, в узлах дерева стоят слова предложения, а ветви помечены именами синтаксических отношений. Всего таких отношений около 80, примерно половину составляют отношения, предложенные в традиционной теории «Смысл ⇔ Текст» И. Мельчука.

Разметка производится автоматически, однако работа над корпусом включает ручную правку всех предложений человеком. Этот этап работы принципиально необходим хотя бы потому, что часто разрешить неоднозначность невозможно без привлечения экстралингвистической информации. Скажем, фраза *Он видел их семью своими глазами* содержит омонимичную словоформу *семью*, однако разрешение этой омонимии, трудное для компьютера, не представляет никаких проблем для человека-постредактора.

Работа по обогащению корпуса семантической информацией включает в себя четыре этапа: (1) разработку инвентаря семантических дескрипторов, (2) создание семантического словаря с приписанными лексемам семантическими дескрипторами и согласование этого словаря с комбинаторным словарем (КС) системы ЭТАП, (3) внедрение семантической информации в уже разме-

ченный морфологически и синтаксически корпус текстов; (4) создание инструментария для работы с семантической информацией.

Два первых этапа работы в настоящее время осуществляет Сектор теоретической семантики ИРЯ РАН, а третий и четвертый – Лаборатория компьютерной лингвистики ИППИ РАН.

Предлагаемый набор семантических дескрипторов (так сказать, семантический метаязык) должен в конечном счете решать две задачи: во-первых, обеспечивать лингвистически содержательную классификацию всей лексики – и предметной, и предикатной, и, во-вторых, в соединении с морфологической и синтаксической разметкой текстов предоставлять исследователю существенную информацию о закономерностях поведения элементов различных лексико-семантических классов в текстах. В качестве дескрипторов везде, где это возможно, используются слова естественного языка, например, слова *занятие* или *деятельность*, в их основных значениях.

При разработке инвентаря семантических дескрипторов мы прежде всего исходим из того, что все лексемы языка делятся на два основных типа – предметные (названия животных, птиц, рыб, овощей, фруктов, камней, гор, планет, светил и т. п.) и предикатные (в данном контексте – любые валентные лексемы). В семантическом метаязыке предусмотрены дескрипторы для обоих классов слов.

Как предметные, так и предикатные дескрипторы делятся на две подгруппы – родовые и видовые. Родовые дескрипторы (genus proximum) обозначаются существительными (например, «животное», «совокупность», «состояние», «действие»), тогда как видовые (differentia speci□ca) – прилагательными (например, «домашний», «природный», «речевой», «ментальный», «физический»). Предикатным словам, кроме родовых и видовых дескрипторов, приписываются семантические роли по каждой из валентностей. Например, глаголу *вязать* в значении «плести спицами или крючком» приписываются семантические роли «агенс»

(***Маша** вяжет*), «результат» (*вязать **шарф***), «пациенс» (*вязать **из шерсти***) и «инструмент» (*вязать **крючком <на спицах>***). С учетом семантических ролей общий объем дескрипторов составляет 250–300 единиц.

Приведем по одному примеру дескрипторного описания предметного и предикатного слова.

АИСТ
DES-OB: «птица», «дикий»

ВЯЗАТЬ
DES-PR: «действие», «физический», «каузация существования»
SEMR1: «агенс»
SEMR2: «результат»
SEMR3: «пациенс»
SEMR4: «инструмент»

Предметной и предикатной лексике соответствуют две разные семантические классификации языковых единиц – таксономическая и фундаментальная.

Предметные дескрипторы членят словарь не с научной, а с наивно-энциклопедической точки зрения. Поэтому, например, слову *паук* будет приписан дескриптор «насекомое» (не «паукообразное»), а элементов научных таксономий типа «хордовые» или «беспозвоночные» вообще не будет. Кроме того, по мере рассмотрения новых предметных лексем в систему могут добавляться и новые дескрипторы.

Сейчас в перечне насчитывается около 90 предметных дескрипторов. Среди них: «вместилище» (для таких слов как *банка, бумажник, ведро, чемодан, шкатулка, ящик*); «знак» (например, для слов *буква, иероглиф, минус* и *цифра*); «прибор» (для слов *барометр, телескоп, часы* и др.); «музыкальный инструмент» (для слов *барабан, пианино, скрипка*); «приспособление» (*замок, капкан, колокол, очки, сеть* и др.); «бытовой» (этот видовой дескриптор приписывается таким словам как *крем, мазь, пылесос* или *щётка*); «природный» (для слов типа *буря, ветер, метель* и

молния); таким словам как *армия, аудитория, банда, бригада* и др. приписывается родовой дескриптор «совокупность» и видовой дескриптор «человеческий».

Предикатные дескрипторы отражают фундаментальную семантическую классификацию предикатов. Используемая нами фундаментальная классификация предикатов разработана Ю.Д. Апресяном[1]. Она существенно отличается от предшествующих классификаций, предложенных в работах Ю.С. Маслова, З. Вендлера, Дж. Лайонза, Т.В. Булыгиной, Е.В. Падучевой[2].

Инвентарь предикатных дескрипторов в нашей системе составляет более 70 родовых и 30 видовых единиц. Мы исходим из того, что список предикатных дескрипторов, в отличие от предметных, должен быть замкнут.

Кроме дескрипторов «верхнего уровня» («действие», «деятельность», «занятие», «воздействие», «свойство», «интерпретация» и др.), в нашей системе выделяются большие подклассы дескрипторов вида «начало» и «прекращение», «каузация» и «ликвидация». Например, дескриптор «начало состояния» приписывается лексемам *мрачнеть, пугаться, слабеть*, «прекращение состояния» – лексемам *забывать* и *опомниться*, «каузация состояния» – *асфальтировать, беспокоить, выпрямлять* и др., «ликвидация состояния» – лексеме *лечить*.

---

[1] *Апресян Ю.Д.* Акциональность и стативность как сокровенные смыслы (охота на *оказывать*) // Сокровенные смыслы: Сборник статей в честь Н.Д. Арутюновой / Гл. ред. Ю.Д. Апресян. М, 2004. С. 13–33; *Апресян Ю.Д.* О семантической непустоте и мотивированности глагольных лексических функций // Вопросы языкознания. 2004. № 4. С. 3–18.

[2] См., например: *Булыгина Т.В.* К построению типологии предикатов в русском языке // Семантические типы предикатов. М., 1982. С. 7–85; *Падучева Е.В.* Семантические исследования: Семантика времени и вида в русском языке; семантика нарратива. М., 1996; *Падучева Е.В.* Динамические модели в семантике лексики. М., 2004.

Использование видовых предикатных дескрипторов («волевой», «качественный», «количественный», «кратный», «однократный», «речевой», «эмоциональный» и мн. др.) позволяет учитывать достаточно тонкие аспекты семантики предикатов.

Так, действия делятся на физические (*ломать*), физиологические (*оправляться*), ментальные (*размышлять*), волевые (*решаться на что-л.*), эмоциональные (обычно каузативные, ср. *злить* во фразах типа *Не зли собаку*), речевые (*требовать*) и социальные (*жениться*). Аналогичные подклассы обнаруживаются и в классах деятельностей, занятий, воздействий и состояний. Так, *дебатировать* обозначает речевую деятельность, а *разглагольствовать* – речевое занятие. *Прогревать* обозначает физическое воздействие, *убеждать* – ментальное (ср. *Даже эти факты его не убедили*), *вынуждать* – волевое (ср. *Это вынуждает меня отказаться от моего намерения*), *удивлять* – эмоциональное (*Его удивил звонок отца*). Аналогичным образом, с некоторыми естественными исключениями, подразделяются состояния: они могут быть физическими (*видеть*), физиологическими (*болен*), ментальными (*знать*), волевыми (*хотеть*), эмоциональными (*бояться*) и социальными (*нуждаться*); речевых состояний, разумеется, нет.

Система семантических ролей, которая используется в корпусе, включает более 50 дескрипторов. Большинство из них вводится в научный оборот впервые в рамках данного проекта.

Среди этих дескрипторов есть как вполне традиционные («агенс», «пациенс», «экспериенсер», «начальная точка» и «конечная точка»), так и новые или получившие новое содержание: «аудитория» для слов типа *отчитываться, оправдываться, рисоваться, щеголять, выпендриваться (**перед кем**)*; «сфера» для описания роли второго актанта предикатов *авторитет (**в науке**), везение (**во всём**)* и т.п. Для случаев расщепления валентности в нашей системе предусмотрены дескрипторы вида «агенс′», «пациенс′» и пр. Так, агенс′ – это часть агенса (рука, нога, глаза и

т.п.), с помощью которой агенс выполняет данное действие: *шевелить пальцами*, *трясти головой*, *вертеть (шляпу в руках)*; пациенс′ – это то в пациенсе, что непосредственно подвергается действию или воздействию: например, *бить (**по спине**), брать (ребенка **за руку**)*; а экспериенсер′ – это «страдающая» часть экспериенсера: *болеть* (ср. *У меня болит **зуб**)*, *мучиться* (***зубами***). У симметричных и некоторых других типов предикатов могут быть повторяющиеся роли, скажем, два агенса или два объекта. В таких случаях для второго из таких актантов вводятся обозначения «агенс2» (например, для слов типа *конфликт (кого **с кем**)*) и «объект2» (например, *Маша – сестра **Люси***).

Предикатные дескрипторы и семантические роли согласованы друг с другом. Так, если предикату приписан в качестве родового дескриптор «действие», «деятельность», «занятие» или «поведение», у его первого актанта будет дескриптор «агенс»; у воздействий первым актантом будет «причина», у процессов – «пациенс», у состояний – «экспериенсер», у свойств – «обладатель». Приведем некоторые примеры.

ВОСПИТАНИЕ
DES-PR: «деятельность», «социальный»
SEMR1: «агенс»
SEMR2: «пациенс»

БАЛОВАТЬСЯ
DES-PR: «поведение», «плохой»
SEMR1: «агенс»

ГРИПП
DES-PR: «состояние», «физиологический», «ненормальный»
SEMR1: «экспериенсер»

ЛЮБОЗНАТЕЛЬНОСТЬ
DES-PR: «свойство», «ментальный»
SEMR1: «обладатель»

На приписывание дескрипторов произвольной лексеме не накладывается никаких формальных ограничений. В частности, одной и той же лексеме может быть приписано несколько дескрипторов одного типа. Например, глаголу *дышать*, у которого есть «ненамеренное» и «намеренное» (ср. *Больной*, д*ышите!)* употребления, будут приписаны дескрипторы «действие» и «процесс». Это же касается глаголов типа *шевелить (пальцами), трясти (головой), греметь (погремушкой), звенеть (уздечками)*). Глаголу *брызгать (водой на стол)* в качестве третьей валентности будут приписаны роли «пациенс» и «место».

Таким образом, все дескрипторы формально трактуются как независимые друг от друга даже в тех случаях, когда между ними есть очевидная семантическая связь. Ср., в дополнение к приведенным выше примерам, класс *везти, вести, водить, возить, гнать, гонять, нести, носить, таскать, тащить* и т. п. с набором дескрипторов «действие», «перемещение», «каузация перемещения»; класс *водить, возить, гонять, носить, таскать* и т. п. с набором дескрипторов «действие», «перемещение», «каузация перемещения», «кратный»; и класс *вести, водить, гнать, гонять, нести, носить, таскать, тащить* и т. п. с набором дескрипторов «действие», «перемещение», «каузация перемещения», «автономный» (в этот класс не войдут глаголы неавтономного перемещения *везти* и *возить*).

Одному слову могут быть одновременно приписаны и предметные, и предикатные дескрипторы. Таковы, например, слова, обозначающие родство:

ОТЕЦ
DES-OB: «человек», «мужской»
DES-PR: «связь», «родственный»
SEMR1: «объект»
SEMR2: «объект2»

Списки дескрипторов могут пересекаться. Так, «время» – это и предметный дескриптор (ср. лексемы *секунда, век, эра*), и пре-

дикатный дескриптор (ср. *длиться*), и семантическая роль (ср. *бежать* (о времени)).

Система дескрипторов устроена так, чтобы по любому дескриптору и любой совокупности дескрипторов из числа приписанных данной лексеме получались лингвистически содержательные классы – лексикографические типы. Так называются классы лексем, у которых есть большое число общих семантически мотивированных несемантических свойств – морфологических, синтаксических, сочетаемостных, коммуникативно-просодических и пр. Выявление закономерностей поведения лексикографических типов в текстах способно дать существенно новое знание о языке.

Приведем некоторые примеры. Все лексемы, которым приписан предметный дескриптор «единица» – *атмосфера, ватт, год, килограмм, километр, рубль, тонна, узел* и т.п. – имеют то общее свойство, что могут входить в количественную группу, реализующую вторую валентность параметрических существительных: *под давлением в сто атмосфер, продолжительностью в два года, со скоростью в 18 узлов* и т.п. Все лексемы, которым приписан предикатный дескриптор «действие», имеют то общее свойство, что способны употребляться в форме императива и подчинять обстоятельство со значением цели.

Мы исходим из предположения, что лексика языка устроена как почти непрерывная сеть, а не строгая иерархия. Поэтому метаязык для ее описания должен обеспечивать разбиение лексики языка на многократно пересекающиеся классы. Так, дескриптор «перемещение» выделяет большой класс лексем типа *бег, бегать1 (по двору), бегать2 (за сигаретами), бегун, бежать, идти, петлять, полет, рыскать, сновать, ходить1 (по комнате), ходить2 (за газетами), ходок, ходьба* и многих других. Дескриптор «занятие» выделяет частично пересекающийся с ним класс глаголов и существительных типа *бегать1, гулять, игра, играть, ходить1, ходьба, читать* (в абсолютивной конструкции) и т.п. Со-

вокупность дескрипторов «занятие», «перемещение», «кратный» и «автономный» позволяет выделить компактный класс лексем (лексикографический тип) *бегать1, бродить, лазать1, летать1, плавать1, ползать1, ходить1* и т.п. с большой совокупностью общих морфологических, синтаксических и сочетаемостных свойств.

Теперь перейдем к описанию технической стороны проекта.

При внесении в корпус семантической разметки должны быть решены следующие задачи:

1) расширение языка разметки, предусматривающее внесение семантической информации;

2) перенос информации из семантического словаря в существующий корпус;

3) расширение функциональности существующего инструментария, позволяющее пользователям работать с семантической информацией.

С самого начала для разметки аннотированного корпуса был выбран язык XML, поскольку он отвечает следующим важным требованиям:

– потенциальная расширяемость языковых конструкций, позволяющая добавлять информацию новых типов;

– наличие стандартных программных средств для разбора языковых конструкций разметки, поиска по размеченному тексту и преобразования разметки.

Используемое подмножество XML позволяет отразить морфо-синтаксическую информацию о слове, такую как имя лексемы (базовая словоформа), морфологические характеристики, имя синтаксического отношения, входящего в слово, идентификатор слова – хозяина синтаксического отношения, и другие типы информации.

Для внесения в корпус семантической информации (предметные и предикатные дескрипторы и семантические роли) в язык будут добавлены конструкции двух типов.

Предметные и предикатные дескрипторы, а также толкования будут описываться как атрибуты нового элемента, описывающего статьи семантического словаря. Для описания связи слов размеченного корпуса с соответствующими статьями этого словаря вводится специальный атрибут.

Приведем пример словарной статьи и отсылки к ней в размеченном корпусе:

<SEM NAME = «ГОВОРИТЬ1» DESPR = «действие, речевой» SEMR1 = «агенс» SEMR2 = «тема» SEMR3 = «содержание» SEMR4 = «адресат» /> <W FEAT = «V НЕСОВ ПРОШ ИЗЪЯВ ЕД МУЖ» ID = «4» LEMMA=«ГОВОРИТЬ» REF = «ГОВОРИТЬ1»> <LINK SYNTR = «релят» DOM = «1» /> говорил </W>

Ролевые дескрипторы слов в корпусе будут записаны в виде ролевых отношений между словом и его семантическими актантами. Ниже приводится пример размеченного предложения *Все, что Мастер говорил о ней, было сущей правдой:*

<S>

<W FEAT = «S ЕД СРЕД ИМ НЕОД» ID = «1» LEMMA = «ВСЕ» REF = «ВСЕ1»> <LINK DOM = «7» SYNTR = «предик» /> ***Все*** *</W>,*

<W FEAT = «S ЕД СРЕД ВИН НЕОД» ID = «2» LEMMA = «ЧТО» REF = «ЧТО1»> <LINK DOM = «4» SYNTR = «1-компл» SEMROLE = «содержание» /> ***что*** *</W>*

<W FEAT = «S ЕД МУЖ ИМ ОД» ID = «3» LEMMA = «МАСТЕР»> <LINK SYNTR = «предик» DOM = «4» SEMROLE = «агенс» / > ***мастер*** *</W>*

<W FEAT = «V НЕСОВ ПРОШ ИЗЪЯВ ЕД МУЖ» ID = «4» LEMMA = «ГОВОРИТЬ» REF = «ГОВОРИТЬ1»> <LINK SYNTR = «релят» DOM = «1» /> ***говорил*** *</W>*

<W FEAT = «PR» ID = «5» LEMMA = «О»> <LINK SYNTR = «3-компл» DOM = «4» SEMROLE = «тема» /> ***о*** *</W>*

<W FEAT = «S ЕД ЖЕН ПР ОД» ID = «6» LEMMA = «ОНА»> <LINK SYNTR = «предл» DOM = «5» /> ***ней*** *</W>,*

<W FEAT = «V НЕСОВ ПРОШ ИЗЪЯВ ЕД СРЕД» ID = «7» REF = «БЫТЬ» LEMMA = «БЫТЬ»> ***было*** *</W>*

<W FEAT = «A ЕД ЖЕН ТВОР» ID = «8» REF = «СУЩИЙ» LEMMA = «СУЩИЙ»> <LINK SYNTR = «опред» DOM = «9» /> ***сущей*** *</W>*

<W FEAT = «S ЕД ЖЕН ТВОР НЕОД» ID = «9» REF = «ПРАВДА1» LEMMA = «ПРАВДА»> <LINK SYNTR = «присвяз» DOM = «7»> ***правдой*** *</W>.*

</S>

Ролевые отношения в первой версии разметки будут указываться, только если соответствующие им семантические актанты параллельны синтаксическим (например, у предикатов типа *купить,* где ролям «агенс» соответствует 1-й столбец модели управления, «пациенс» – 2-й столбец и т.д.; ср. *Иван купил корову*). Для таких конструкций как *красный шар* (семантическая валентность слова *красный* выражается синтаксически подчиняющим его словом *шар)* или *он может работать* (семантическая валентность слова *работать* выражается словом *он,* синтаксически зависящим от модального глагола *может)*, ролевые отношения будут опущены.

Перенос семантической информации в корпус будет заключаться во внесении в размеченный текст ссылок на статьи семантического словаря и приписывании словам текста информации о семантических ролях, которые эти слова реализуют. Большая часть этих работ может быть автоматизирована. Вмешательство человека потребуется:

– при внесении ссылок на статьи семантического словаря для омонимичных слов из корпуса, если омонимы невозможно различить по части речи;

– при проверке результатов разметки.

Для поддержки работы с семантической информацией, добавляемой в корпус, инструментарий должен обеспечивать:

– внесение семантической информации при предварительной автоматической разметке текстов, добавляемых в корпус;

– просмотр и редактирование семантического словаря, ссылок на статьи семантического словаря и семантических ролей.

Поскольку инструментарий для работы с корпусом в качестве источника лингвистических данных использует КС ЭТАПа, семантическая информация должна быть частично перенесена из семантического словаря в КС. В первую очередь это касается ролевых дескрипторов, для которых соответствующие семантические актанты параллельны синтаксическим. Они будут приписаны соответствующим столбцам модели управления. Также в статье КС будет указана ссылка на соответствующую статью семантического словаря.

Перенос семантической информации в КС может быть в значительной степени осуществлен автоматически, однако в некоторых случаях может потребоваться вмешательство человека. В первую очередь это касается установления соответствия между разбиением значений неоднозначных слов в словарях. Также не всегда возможно автоматическое сопоставление ролевых дескрипторов и столбцов моделей управления.

Из оставшихся задач только просмотр и редактирование семантического словаря потребует введения нового функционала. Остальные могут быть решены за счет незначительного изменения существующих.

Внешний вид пользовательского интерфейса для редактирования семантического словаря и для просмотра семантической информации в предложении приводится на рис. 2–3.

Таким образом, внесение семантической разметки потребует не слишком больших трудозатрат. Первого варианта разметки можно ожидать к концу следующего года.

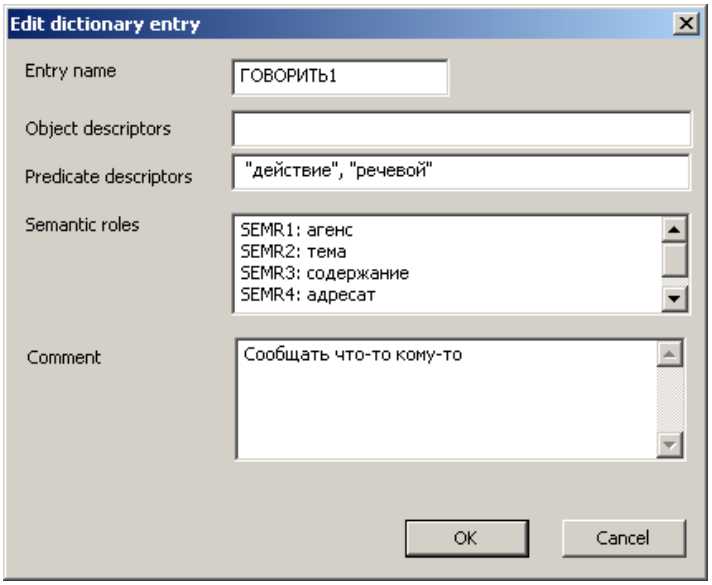

Рис. 1. Вид пользовательского интерфейса
для редактирования семантического словаря

Рис. 3. Вид пользовательского интерфейса
для просмотра семантической информации в предложении