Ю. Д. Апресян[1,2], И. М. Богуславский[1,3], Б. Л. Иомдин[2], Л. Л. Иомдин[1], А. В. Санников[2], В. З. Санников[1], В. Г. Сизов[1], Л. Л. Цинман[1]

СИНТАКСИЧЕСКИ И СЕМАНТИЧЕСКИ АННОТИРОВАННЫЙ КОРПУС РУССКОГО ЯЗЫКА: СОВРЕМЕННОЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ*

[1]Институт проблем передачи информации РАН, Москва, Россия
[2]Институт русского языка им. В. В. Виноградова РАН, Москва, Россия
[3]Политехнический университет, Мадрид, Испания

Вводные замечания

В работе излагается современное состояние глубоко аннотированного корпуса русских текстов, который в течение последних пяти лет разрабатывается Лабораторией компьютерной лингвистики ИППИ РАН в сотрудничестве с Сектором теоретической семантики ИРЯ РАН [Boguslavsky *et al.* 2000, Богуславский *и др.* 2002, 2003; Апресян *и др.* 2004].

Описываемый корпус основан на идеологии разработанного в ИППИ РАН многоцелевого лингвистического процессора ЭТАП (см. о нем, в частности, Апресян *и др.* 1989, 1992, Apresjan *et al.* 2003) и в первую очередь русского синтаксического анализатора системы машинного перевода ЭТАП-3, которая является центральной частью этого процессора. Корпус является составной, но полностью автономной частью исследовательского проекта общероссийского масштаба – Национального корпуса русского языка (www.ruscorpora.ru), которому посвящена значительная часть работ настоящего сборника.

Главное достоинство описываемого корпуса – глубина аннотации текста: каждое его предложение снабжено полной морфологической разметкой со снятой омонимией и полной синтаксической структурой. В настоящее время наш корпус – единственный лин-

* Данная работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (гранты № 05-04-04190а и 04-04-00263а) и Российского фонда фундаментальных исследований (грант № 04-07-90179). Обоим фондам авторы выражают глубокую признательность.

гвистический ресурс для русского языка, содержащий синтаксическую разметку. Корпус постоянно растет: в настоящее время его объем приближается к 25 000 синтаксически размеченных предложений, или свыше 350 000 словоупотреблений[1]. Тексты, используемые в корпусе, относятся к разнообразным литературным жанрам (художественная проза XX века, современная научно-популярная литература, публицистика, газетные и журнальные материалы, новостные ленты). В настоящее время доступ к синтаксически аннотированному корпусу русского языка открыт в тестовом режиме на сайте Национального корпуса русского языка по адресу http://-www.ruscorpora.ru/search-syntax.html.

В последние два года создатели корпуса ведут активную работу над обогащением синтаксически аннотированного корпуса семантической разметкой.

Принципы синтаксического аннотирования корпуса излагаются в первой части настоящей работы, а принципы семантического аннотирования – во второй части.

1. Синтаксическое аннотирование

Глубоко аннотированный корпус русского языка организован следующим образом. Каждый входящий в него текст представляет собой отдельный файл в xml-формате, который содержит морфологическую информацию обо всех словоформах данного текста (т. е. имя лексемы, соответствующей данной словоформе, и набор ее грамматических характеристик), а также синтаксическую структуру (СинтС) каждого предложения в виде дерева зависимостей.

Рассмотрим пример. На рис. 1 представлена СинтС предложения

(1) *Наибольшее возмущение участников митинга вызвал продолжающийся рост цен на бензин, устанавливаемых нефтяными компаниями*

Как видно из рисунка, в узлах дерева зависимостей СинтС стоят слова предложения (1), представленные именами лексем (в прямоугольниках) и цепочками грамматических характеристик (справа от прямоугольников), а ветви помечены именами синтаксических отношений (в овалах). Всего используется около 80 таких отношений,

[1] Помимо авторов статьи, в работе по составлению синтаксически аннотированного корпуса активное участие принимали сотрудники Лаборатории компьютерной лингвистики ИППИ РАН С. А. Григорьева, Л. Г. Крейдлин, А. В. Лазурский, Л. Г. Митюшин.

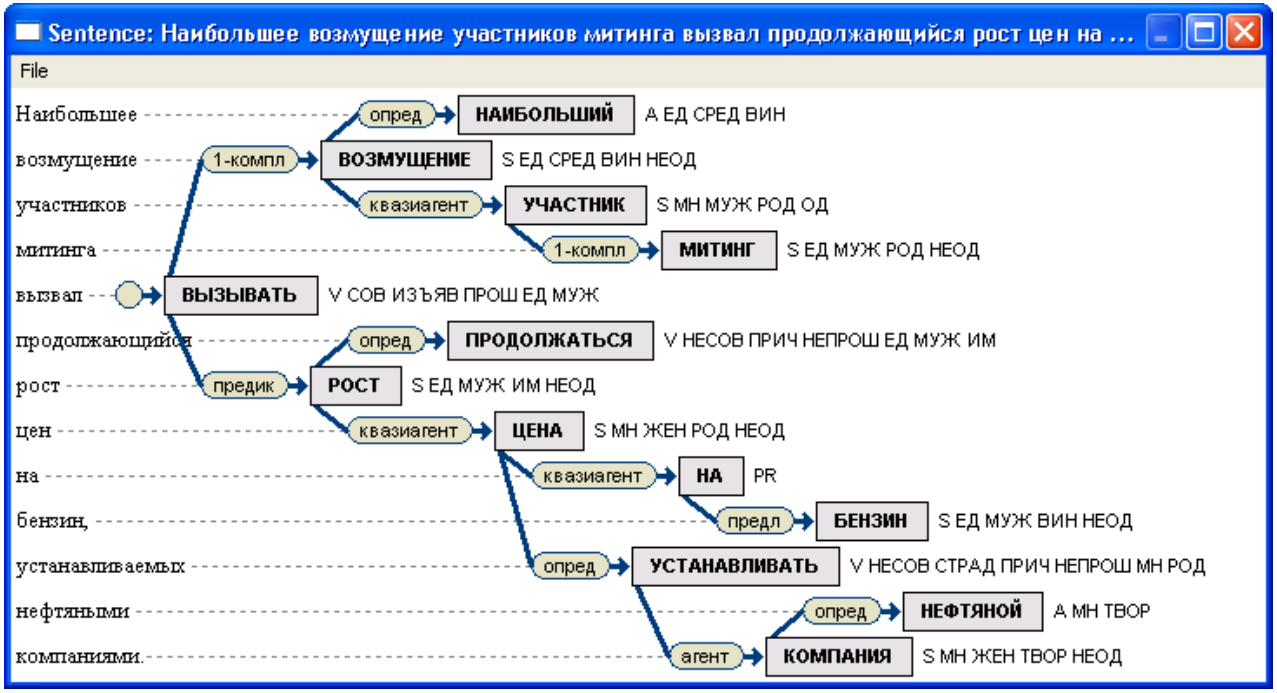

Рис. 1. Синтаксически размеченное предложение

примерно половину из которых составляют отношения, предложенные в традиционной теории «Смысл ⇔ Текст» И. А. Мельчука.

Корпус текстов строится полуавтоматически: каждое предложение пропускается через морфологический и синтаксический анализаторы модуля русско-английского перевода системы ЭТАП-3, после чего полученная СинтС проверяется и при необходимости корректируется редакторами-лингвистами. Анализаторы пользуются всеми лингвистическими ресурсами системы ЭТАП-3 – автоматическим морфологическим словарем русского языка объемом в 120000 лексических единиц, комбинаторным словарем объемом около 90000 лексических единиц и достаточно полной грамматикой русского языка, включающей сотни синтаксических правил (синтагм).

Для удобства работы редактора используется специальный программный комплекс, состоящий из модуля разбивки текста на фразы и мощного графического редактора структур, способного работать с xml-файлами (стандартный формат для лингвистических ресурсов подобного рода) и позволяющего легко и быстро модифицировать древесные объекты. Новейшая версия этого программного комплекса дополнительно оптимизирует работу редактора, обеспечивая контроль в реальном времени за вносимыми исправлениями (полнотой и непротиворечивостью набора грамматических характеристик, допустимостью фрагментов дерева зависимостей и т. п.) и исправляя «на лету» значительную часть редакторских ошибок и опечаток.

Этап постредактирования синтаксических структур до помещения их в корпус принципиально необходим хотя бы потому, что содержащиеся в текстах предложения часто неоднозначны, причем во многих ситуациях разрешить неоднозначность невозможно без привлечения экстралингвистической информации, в принципе недоступной автоматическому анализатору. Скажем, простая фраза *Он увидел их семью своими глазами,* хотя и содержит омонимичную словоформу *семью* (вин. п. существительного СЕМЬЯ и твор. п. числительного СЕМЬ), не представляет никаких проблем для человека, разрешающего эту неоднозначность, но может создать непреодолимые трудности для компьютерного анализа.

Следует отметить, что СинтС предложений, входящих в синтаксически размеченный корпус, редактируются весьма тщательно[2]. Специально для корпуса был разработан ряд решений относительно формы представления СинтС, не использующихся в анализаторах ЭТАП-3, но необходимых для лучшего восприятия корпуса лингвистами. Среди этих решений особого упоминания заслуживают два.

1) Восстановление некоторых типов синтаксического эллипсиса. В случаях, если опущенные слова физически присутствуют в другой части предложения (например, в конструкциях с так называемым сочинительным сокращением), эти слова восстанавливаются. Например, в СинтС предложения

(2) *Я купил чемодан, а он сумку*

между *он* и *сумку* вставляется узел «ПОКУПАТЬ» с соответствующим набором характеристик и пустым текстовым элементом. От этих «фантомных» слов проводятся все необходимые связи (см. рис. 2):

При отсутствии такого восстановленного узла древесная СинтС предложения (2) выглядела бы весьма неестественно. На рис. 3 показана непрепарированная СинтС, построенная для (2) синтаксическим анализатором системы ЭТАП-3.

При восстановлении эллипсиса леммы во вновь введенных словах совпадают с теми, которые уже встретились в предложении, а отдельные морфологические характеристики могут меняться (так, в предложении *Я купил чемодан, а она сумку* характеристика муж в новом, «фантомном» глагольном узле ПОКУПАТЬ заменяется на жен).

2) Использование специальных «фиктивных» слов для частичной нормализации высокоэллиптичных предложений. Этот прием используется в случаях, когда в предложении «опущен» глагол некоторой размытой семантики, как в предложениях текста

(3) *Парочку морей бы еще в Сибирь. Африку можно бы ниже. Индия пусть.* (Т. Толстая).

Во всех таких случаях добавляется глагольный узел, ему приписываются наиболее естественные грамматическими характеристики,

[2] Это, впрочем, не исключает того, что, несмотря на все усилия редакторов СинтС, в корпусе могут содержаться ошибки.

Рис.2. Синтаксическая структура предложения с восстановленным эллипсисом

Рис.3. Непрепарированная синтаксическая структура эллиптичного предложения

а в качестве леммы пишется НЕОПР-ГЛАГОЛ (неопределенный глагол) и затем в скобках глагол, который является «естественной гипотезой». Так, в первом предложении текста (3) после слова *еще* добавляется узел с леммой НЕОПР-ГЛАГОЛ (ДОБАВИТЬ), во втором предложении после слова *бы* – НЕОПР-ГЛАГОЛ (ОПУСТИТЬ), а в последнем предложении после *пусть* – узел с леммой НЕОПР-ГЛАГОЛ (ОСТАВАТЬСЯ).

На рис. 4 (с. 200) показано, как в СинтС предложения

(4) *Мы надавим на министра, а вы от дополнительной суммы, полученной из федерального бюджета, — едва ли не подмигнул депутат, — заказ нашей фирме*

вставлен узел с леммой НЕОПР-ГЛАГОЛ (ДЕЛАТЬ).

Обратим также внимание на то, что в СинтС предложений корпуса некоторые слова, пишущиеся в тексте раздельно, представляются в виде цельных лексем (как в случае с *едва ли* на рис. 4, в случаях безусловных оборотов типа *во что бы то ни стало* и т. д.); с другой стороны, в некоторых случаях слова, фигурирующие в тексте как цельные единицы, приходится разделять, как в предложении

(5) *Где-то он теперь живет?*,

в котором имеет место не неопределенное местоименное наречие *где-то* ‘в некотором месте’, а вопросительное местоименное наречие *где,* за которым следует частица *то:*

Рис.5. Синтаксическая структура предложения с узлом, выделенным из единой текстовой единицы

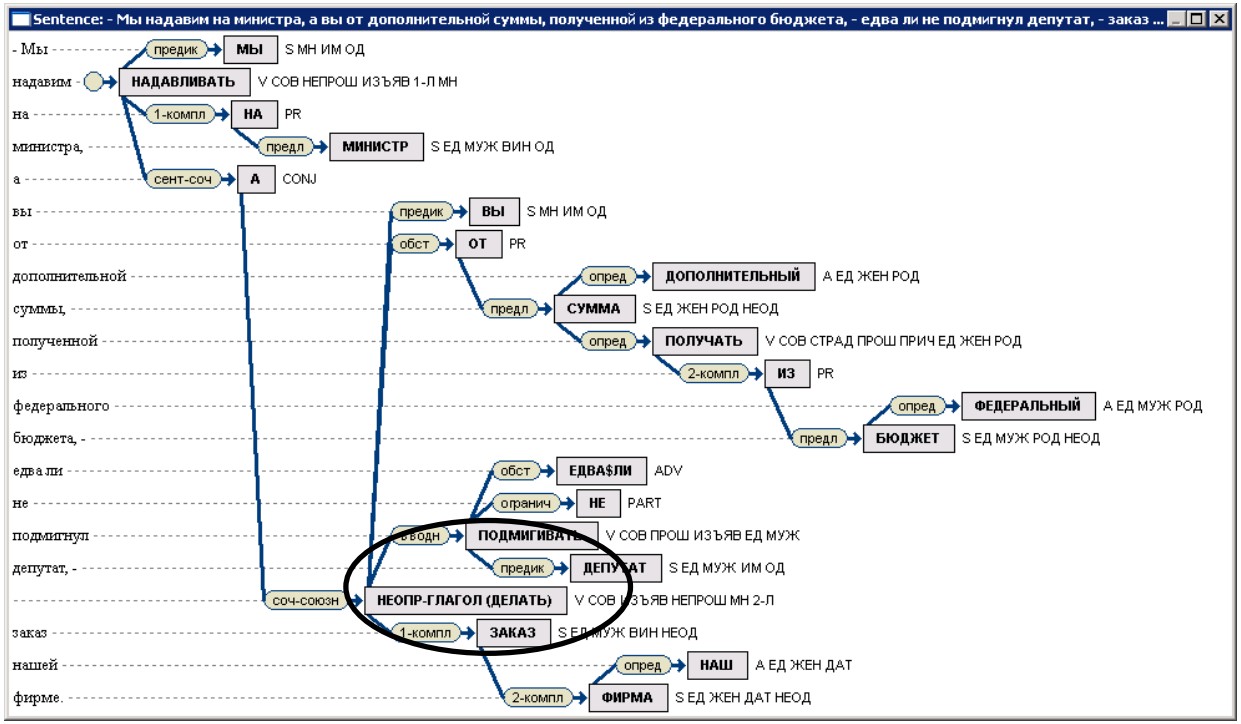

Рис.4. Синтаксическая структура с «фантомным» глаголом размытой семантики

С помощью синтаксически аннотированного корпуса можно решать самые разные лингвистические задачи. В частности, синтаксически размеченные корпусы широко используются для создания компьютерных программ, автоматически извлекающих из текстов разные виды информации. Обзор этих задач выходит за рамки настоящей статьи. Мы кратко остановимся лишь на использовании корпуса в качестве источника обратной связи для развития синтаксического анализатора системы ЭТАП-3, с помощью которого он и был порожден (подробнее об этом см. Богуславский *и др.* 2003).

Не будет преувеличением сказать, что проблема разрешения языковой неоднозначности по сложности занимает первое место среди всех задач автоматической обработки текстов. Системы автоматического перевода не являются исключением: сколь бы тщательно ни были разработаны грамматики и словари такой системы, ее анализирующие модули на любом этапе сталкиваются с необходимостью выбора вариантов анализа текстового материала – будь то на уровне морфологии, синтаксиса или семантики.

Неудовлетворительная ситуация с разрешением неоднозначности в системах автоматической обработки текста носит универсальный характер и, вообще говоря, не зависит от лингвистической модели, лежащей в основе анализатора. Это и естественно: то, что в процессе понимания текста с легкостью делает человек, опирающийся при выборе интерпретации на здравый смысл, знания о мире и широкий контекст коммуникации, пока недоступно никаким компьютерным системам. Представления о мироустройстве, по-видимому, вообще не поддаются сколько-нибудь масштабной формализации. Да и степень формализации чисто языковой семантики, достигнутая в компьютерной лингвистике, пока не достаточна для того, чтобы эксплицировать те нетривиальные сведения о смысле высказывания, которыми необходимо располагать для разрешения неоднозначности текста.

Это, однако, не означает, что попытки решить проблему неоднозначности при автоматическом переводе вообще лишены перспективы. Частичное решение этой проблемы вполне возможно, и всякая система анализа естественного языка располагает арсеналом более или менее действенных средств, направленных на сокращение неоднозначности в ходе обработки текста, – от простого

игнорирования редких лексических единиц или синтаксических конструкций до использования масштабных статистических процедур, определяющих частотность встречаемости отдельных языковых элементов.

В ходе развития лингвистического процессора ЭТАП-3 его авторы в сотрудничестве с другими коллегами теоретически разрабатывали гибридную стратегию, сочетающую детерминистский (опирающийся на правила) и статистический подход к анализу текста [Carl *et al.* 2000, Streiter *et al.* 2000a], а также предпринимали попытки практически реализовать эту стратегию. В частности, был предложен статистический способ использования двуязычных корпусов текстов для оптимизации выбора переводных эквивалентов словосочетаний [Streiter *et al.* 2000b] и разработана система статистически обоснованных приоритетов, динамически приписываемых элементам строящейся структуры предложения на разных этапах синтаксического анализа и ориентирующая анализатор на построение оптимальной структуры [Иомдин *и др.* 2001, Iomdin *et al.* 2002].

С появлением синтаксически размеченного корпуса у авторов системы возникла идея использовать собранную на его основе статистическую информацию для оптимизации алгоритма синтаксического анализа. Суть этого метода такова. На начальной стадии работы блока синтаксического анализа его правила (синтагмы) порождают множество минимальных поддеревьев (два узла, связанных синтаксическим отношением) – своего рода кирпичиков, из которых будет построено все дерево зависимостей. Затем осуществляется выбор вершины будущего дерева зависимостей и его непосредственное построение. На этой стадии происходит частичное разрешение синтаксической неоднозначности (иными словами, удаление части минимальных поддеревьев). Здесь используется система различных фильтров, центральную роль в которой играет механизм приоритетов. В этот момент в действие вступает новый статистический блок, играющий роль дополнительного фильтра. Основные идеи, легшие в основу его создания, были предложены в работах [Чардин 2001, 2003]. Статистический блок взвешивает минимальные поддеревья, а также цепочки минимальных поддеревьев длиной в три слова в пространстве поиска алгоритма синтаксического анализа на основании частоты встречаемости фрагментов такого вида в деревьях зависимостей корпуса. В итоге в системе

формируются новые значения приоритетов связей, которые вычисляются с учетом вновь приписанных весов.

Серия экспериментов с участием нового блока показала, что в определенных ситуациях синтаксические структуры, произведенные по-новому, отличаются от результатов работы стандартного алгоритма ЭТАПа-3, и применение корпусно-статистического модуля дает положительный эффект. Приведем пример. Для предложения

(6) *В экстремальных условиях прошли в Ленинградской области гонки седьмого этапа розыгрыша Кубка мира по лыжному спорту*

СинтС, построенная с использованием корпусной статистики, имела вид такой, какой показан на рис. 6 (с. 204).

Если отвлечься от неточности в присоединении предложной группы *по лыжному спорту* (которую следовало бы подчинить не слову *гонки*, а слову *кубок*), эта СинтС (1′) была построена правильно. В частности, верно была определена синтаксическая роль существительного *гонки* как подлежащего при вершине предложения *прошли*. При отключении статистического модуля алгоритм построил для (6) СинтС, в которой слово *гонки* оказалось зависящим от своего непосредственного соседа *области*, а вершина *прошли* осталась без подлежащего. Поскольку структура эта явно ошибочна, совершенно неадекватным получился и построенный на ее основе английский перевод.

Несмотря на ограниченный объем экспериментов с использованием корпусной статистики, в целом такой подход следует признать перспективным.

2. Семантическая разметка

Работа по обогащению корпуса семантической информацией (см. о ней, в частности, Апресян *и др.* 2004) 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 системы ЭТАП, (3) внедрение семантической информации в уже размеченный морфологически и синтаксически корпус текстов; (4) создание инструментария для работы с семантической информацией.

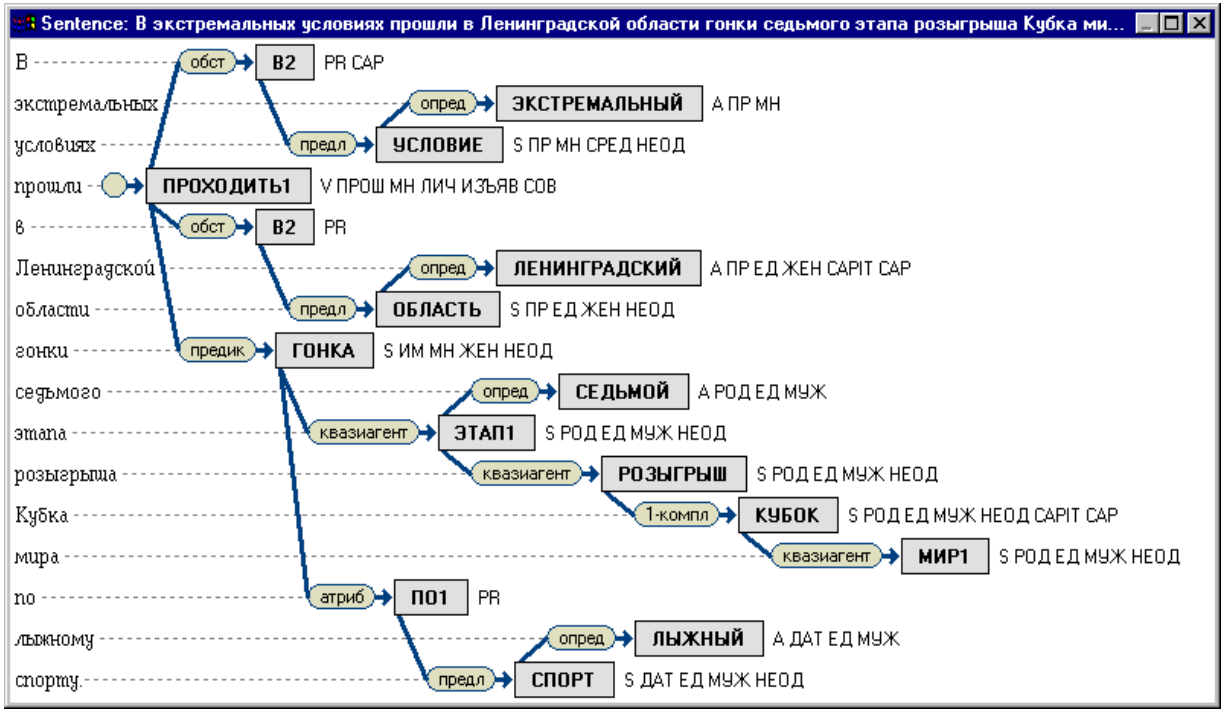

Рис. 6. СинтС, построенная с использованием корпусной статистики

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 (такие, например, как слова *занятие* или *деятельность*) в их основных значениях. В некоторых случаях приходится, однако, использовать лингвистические термины типа *каузация существования*.

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 (упрощенно говоря, любые валентные лексемы). В семантическом метаязыке предусмотрены дескрипторы для обоих классов слов.

Как предметные, так и предикатные дескрипторы подразделяются на две подгруппы – родовые и видовые. Родовые дескрипторы (genus proximum) обозначаются существительными (например, ‘животное’, ‘совокупность’, ‘состояние’, ‘действие’), тогда как видовые (differentia specifica) – 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’ (*вяжет* ***шарф***), ‘пациенс’ (*вяжет* ***из шерсти***) и ‘инструмент’ (*вяжет* ***крючком*** ***<на спицах>***). С учетом семантических ролей общий объем дескрипторов составляет около 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* приписывается дескриптор ‘насекомое’ (а не ‘паукообразное’), а элементов научных таксономий типа ‘хордовые’ или ‘беспозвоночные’ вообще не предусматривается. 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 [Апресян 2004а, 2004б]. Она существенно отличается от предшествующих классификаций, предложенных в работах Ю. С. Маслова, З. Вендлера, Дж. Лайонза, Т. В. Булыгиной, Е .В. Падучевой и других исследователей (см. в частности [Булыгина 1982; Падучева 1996, 2004]).

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 нашем семантическом языке 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’ – лексемам *лечить, чинить.*

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 разработаны специально для целей семантической разметки корпуса и вводятся в научный оборот впервые.

Среди этих дескрипторов есть как вполне традиционные (‘агенс’, ‘пациенс’, ‘экспериенсер’, ‘начальная точка’ и ‘конечная точка’), так и новые или получившие новое содержание: ‘аудито-

рия’ 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**)*; 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 еще несколько примеров.

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, од-

ной и той же лексеме может быть приписано несколько дескрипторов одного типа. Например, глаголу *дышать*, у которого есть «ненамеренное» (*Иван запыхался и тяжело дышал*) 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, эпоха*), и предикатный дескриптор (ср. *длиться, медлить*), 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, синтаксиче-

ских, 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, (морской) 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 убеждения, 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, бегать 1 (по двору), бегать 2 (за сигаретами), бегун, бежать, идти, петлять, полет, рыскать, сновать, ходить 1 (по комнате), ходить 2 (за газетами), ходок, ходьба* и многих других. Дескриптор ‘занятие’ выделяет частично пересекающийся с ним класс глаголов и существительных типа *бегать 1, гулять, игра, играть, ходить 1, ходьба, читать* (в абсолютивной конструкции) и т. п. Совокупность дескрипторов ‘занятие’, ‘перемещение’, ‘кратный’ и ‘автономный’ позволяет выделить компактный класс лексем (лексикографический тип) *бегать 1, бродить, лазать 1, летать 1, плавать 1, ползать 1, ходить 1* и т. п. с большой совокупностью общих морфологических, синтаксических и сочетаемостных свойств.

После создания инвентаря собственных и ролевых дескрипторов авторами был составлен семантический словарь объемом приблизительно в 3000 единиц, в котором каждой лексеме сопоставлены семантические дескрипторы и роли[3]

[3] Помимо авторов статьи, в разработке семантического языка и составлении экспериментального семантического словаря участвовали сотрудники Института русского языка РАН В. Ю. Апресян, О. Ю. Богуславская, Т. В. Крылова, И. Б. Левонтина и Е. В. Урысон.

В настоящее время производится пробная семантическая разметка фрагмента существующего синтаксического корпуса.

На рис. 7 (с. 212) приводится семантическая разметка для предложения 1.

В квадратных скобках справа от лексической единицы приведены собственные дескрипторы соответствующих слов, а ветви вместо имен синтаксических отношений помечены именами семантических ролей.

В первой версии семантической разметки предполагается указывать ролевые отношения только тогда, когда соответствующие им семантические актанты параллельны синтаксическим (например, у предикатов типа *покупать,* где роли 'агенс' соответствует первый столбец модели управления, роли 'пациенс' – второй 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.

Литература

Апресян 2004а – *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.: Языки русской культуры, 2004, с. 13-33.

Апресян 2004б – *Апресян Ю. Д.* О семантической непустоте и мотивированности глагольных лексических функций // Вопросы языкознания, № 4, 2004, с. 3-18.

Апресян *и др.* 1989 – *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Перцов Н. В., Санников В. З., Цинман Л. Л.* Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.

Апресян *и др.* 1992 – *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Санников В. З., Цинман Л. Л.* Лингвистический процессор для сложных информационных систем. М.: Наука, 1992.

Апресян *и др.* 2004 – *Апресян Ю. Д., Иомдин Л. Л., Санников А. В., Сизов В. Г.* Семантическая разметка в глубоко аннотированном корпусе русского языка // Труды международной конференции «Корпусная лингвистика – 2004». СПб: Изд-во Санкт-Петербургского университета, 2004, с. 41-54.

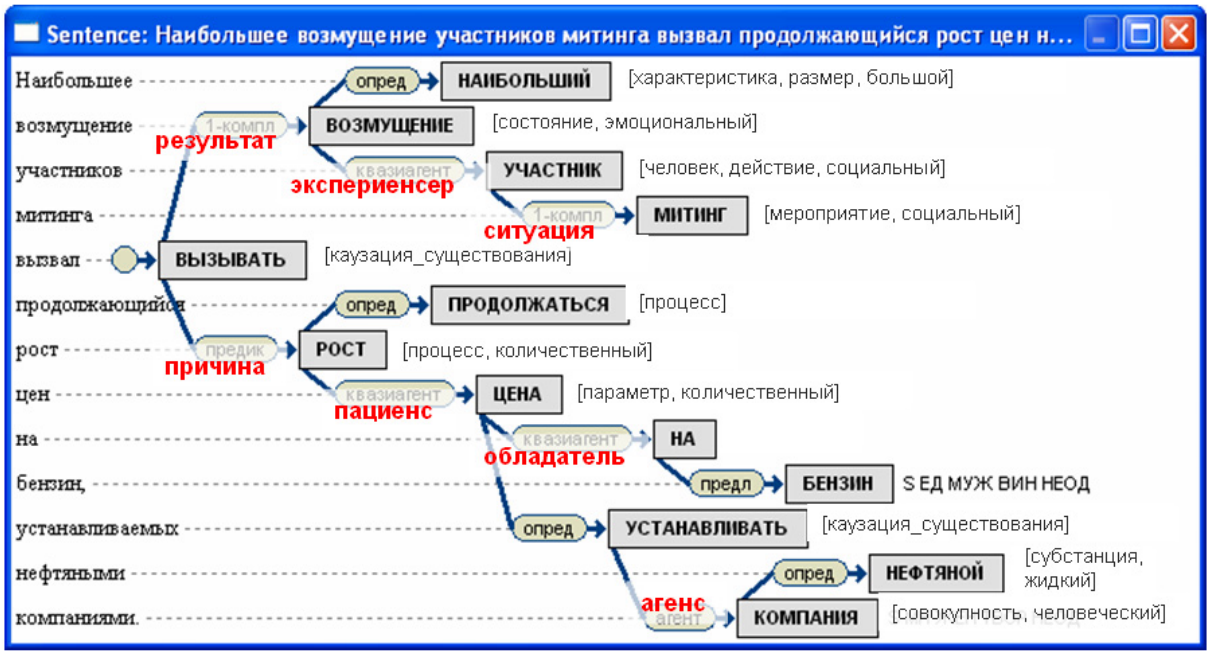

Рис. 7. Семантически размеченное предложение

Богуславский *и др.* 2002 – *Богуславский И. М., Григорьев Н. В., Иомдин Л. Л., Крейдлин Г. Е., Фрид Н. Е.* Разработка синтаксически размеченного корпуса русского языка // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб: Изд-во Санкт-Петербургского университета, 2002, с. 40-50.

Богуславский *и др.* 2003 – *Богуславский И. М., Иомдин Л. Л., Сизов В. Г., Чардин И. С.* Использование размеченного корпуса текстов при автоматическом синтаксическом анализе // Cognitive modeling in linguistics-2003. Варна, 2003.

Булыгина 1982 – *Булыгина Т. В.* К построению типологии предикатов в русском языке // Семантические типы предикатов. М., 1982, с. 7-85.

Иомдин *и др.* 2001 – *Иомдин Л. Л., Сизов В. Г., Цинман Л. Л.* Использование эмпирических весов при синтаксическом анализе // Обработка текста и когнитивные технологии. Труды конференции «Когнитивное моделирование». Казань: Отечество, № 6, 2001, с. 64-72.

Падучева 1996 – *Падучева Е. В.* Семантические исследования (Семантика времени и вида в русском языке; семантика нарратива). М.: Школа «Языки русской культуры», 1996.

Падучева 2004 – *Падучева Е. В.* Динамические модели в семантике лексики. М.: «Языки русской культуры», 2004.

Чардин 2001 – *Чардин И. С.* Использование аннотированного корпуса для снятия синтаксической неоднозначности в лингвистическом процессоре ЭТАП-3 // Материалы 2-ой Всероссийской конференции «Теория и практика речевых исследований» (АРСО-2001). М.: Изд-во МГУ, 2001, с.26-27.

Чардин 2003 – *Чардин И. С.* Лингвистические корпуса с синтаксической разметкой и их применение // Научно-техническая информация, 2003, №5.

Apresjan *et al* 2003 — *Jurij Apresian, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Vladimir Sannikov, Victor Sizov, Leonid Tsinman.* ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. // MTT 2003, First International Conference on Meaning — Text Theory. Paris, Ecole Normale Supérieure, Paris, June 16-18 2003, pp. 279-288.

Boguslavsky *et al.* 2000 — *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, 987-991. Saarbrücken, 2000.

Carl et al. 2000 — *Carl M., Pease C., Streiter O., Iomdin L.* Towards a Dynamic Linkage of Example-Based and Rule-Based Machine Translation // Machine Translation, 2000.

Iomdin *et al.* 2002 — *Iomdin L., Sizov V., Tsinman L.* Utilisation des poids empiriques dans l'analyse syntaxique: une application en traduction automatique // META, vol. 47, No. 3, 2002, p. 351-358.

Streiter *et al.* 2000a — *Streiter O., Iomdin L., Sagalova I.* Learning Lessons from Bilingual Corpora: Benefits for Machine Translation // International Journal of Corpus Linguistics. Vol. 5(2), 2000, pp. 199-230.

Streiter *et al.* 2000b — *Streiter O., Iomdin L., Carl M.* A Virtual Machine for Hybrid Machine Translation // Труды Международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям. Т. 2. Протвино, 2000, с. 382-393.