Титульный лист отчета о работе в 2012 г.
по **Программе фундаментальных исследований Президиума РАН «Корпусная лингвистика»**

<table>
<tr><td colspan="2">Номер и название направления Программы 1. Создание и развитие корпусных ресурсов по современному русскому языку</td></tr>
<tr><td colspan="2">Название проекта ФреймБанк: разметка семантических ролей и морфосинтаксического оформления участников фреймов (на базе НКРЯ)</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Падучева Елена Викторовна, д. филол. н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>elena.paducheva@yandex.ru,
olesar@gmail.com,</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Молдован Александр Михайлович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Глебова Татьяна Николаевна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
тел.: (495) 695 26 60, факс: (495) 695 26 03;
ruslang@ruslang.ru</td></tr>
<tr><td>Год начала – год окончания проекта</td><td>2012—2014</td></tr>
<tr><td>Объем финансирования, полученного в 2012 г.</td><td>Объем финансирования, запрашиваемый на 2013 г.</td></tr>
<tr><td>Исполнители (ФИО, уч. ст.)</td><td>Ляшевская О.Н., к.ф.н., ИРЯ РАН
Добрушина Е.Р., к.ф.н., ПСТГУ
Кустова Г.И., д.ф.н., МПГУ
Бонч-Осмоловская А.А., к.ф.н., НИУ ВШЭ
Толдова С.Ю., к.ф.н., РГГУ
Резникова Т.И., к.ф.н., НИУ ВШЭ
Исаев Д.Ю., НИУ ВШЭ
Кудинов М.С., МГУ
Кашкин Е.В., МГУ
Митрофанова О.А., к.ф.н., СПбГУ
Шиморина А.С., ИЛИ РАН</td></tr>
<tr><td>Дата сдачи отчета
20.11.2012</td><td>Подпись руководителя проекта:</td></tr>
</table>

Форма 1

УТВЕРЖДАЕМ

Координатор Программы акад. Вяч. Вс. Иванов

Координатор Программы чл-корр. РАН В.А. Плунгян

«___»_____2012 г.

1. Название направления **1. Создание и развитие корпусных ресурсов по современному русскому языку**

2. Название проекта **ФреймБанк: разметка семантических ролей и морфо-синтаксического оформления участников фреймов (на базе НКРЯ)**

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы) **Падучева Елена Викторовна, д. филол. н., проф., г.н.с. ВИНИТИ РАН**

4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)

Ляшевская Ольга Николаевна, к.ф.н., с.н.с. ИРЯ РАН, Добрушина Екатерина Роальдовна, к.ф.н., доцент ПСТГУ, Кустова Галина И., д.ф.н., профессор МПГУ; Бонч-Осмоловская Анастасия Александровна, к.ф.н., проф. НИУ ВШЭ; Толдова Светлана Юрьевна, к.ф.н., доцент РГГУ; Резникова Татьяна Исидоровна, к.ф.н., доцент НИУ ВШЭ; Исаев Дмитрий Юрьевич, магистрант НИУ ВШЭ; Кудинов Михаил Сергеевич, студент МГУ; Кашкин Егор Владимирович, аспирант МГУ; Митрофанова Ольга Александровна, к.ф.н., доцент СПбГУ; Шиморина Анастасия Сергеевна, аспирант ИЛИ РАН

5. Содержание фактически проделанной работы, полученные результаты, их новизна, научная и практическая значимость, актуальность, соответствие мировому научному уровню (до 5 стр.)

В 2012 году работа велась по двум направлениям: 1) разметка реализаций аргументно-предикатных и идиоматических конструкций в примерах письменной речи НКРЯ и пополнение словаря конструкций системы; 2) разработка иерархического инвентаря семантических ролей (экспликаций).

5.1. Разметка текстовых реализаций лексических конструкций на корпусе данных является новой задачей в русской корпусной лингвистике. Каждый пример соотносится с соответствующей конструкцией в словаре, аннотиру-

ется выражение/невыражение участника, синтаксические особенности кодирования (ранг, связь с предикатом через контроль, изменение ранга в связи с пассивной, аттрибутивной причастной и др. типами конструкций и т.п.), особенности предложно-падежного и др. морфосинтаксических способов оформления участников; также аннотируются сирконстантные участники ситуации и модальные элементы.

В 2012 г. система ФреймБанк была пополнена представительными выборками примеров из НКРЯ (около 100 примеров на глагол) для 500 глаголов и предикатных имен. Отредактирована разметка корпусных данных для 1000 ранее размеченных лексических единиц.

Данный тип разметки представляет собой семантически более глубокую версию корпусной разметки, нежели синтаксическая разметка трибанков. Присутствующая в корпусе разметка лексико-семантических классов, к которым принадлежат леммы, пополняется разметкой синтагматических отношений между разными элементами текста. Так называемая ФреймНет-разметка имеет большую значимость как для академических исследований семантико-синтаксического интерфейса, так и для работ с использованием фреймовой разметки – data mining, opinion mining, лексикография и т.д.

Наиболее близкие аналоги в мировой корпусной лингвистике – PDT (Пражский глубоко размеченный корпус); FrameNet (имеет менее глубокий уровень разметки и отслеживания связей между примерами и словарем).

5.2. В 2012 г. для системы Framebank был разработан инвентарь экспликаций (семантических ролей), в соответствии с которым была произведена разметка шаблонов конструкций в словаре системы (более 10 000 конструкций). Инвентарь представляет собой иерархию (граф), включающий ядерные элементы (основные семантические роли) и более периферийные, частные экспликации, соответствующие узкому тематическому классу, семантическому ряду или даже индивидуальным предикатам. Иерархия ролей строилась на следующих принципах:

1) инвентарь ролей может быть представлен на нескольких уровнях дробности, от краткого списка ядерных ролей (порядка 10) до детального списка очень частных экспликаций (от 200 до 1000 и более экспликаций);

2) каждая частная экспликация сводится к одной или (реже) нескольким ядерным ролям; представляя их периферийный случай. Например, созданная в стиле FrameNet экспликация для участника фрейма *Вася поет* «тот, кто поет», возводится к экспликации «говорящий» (*петь* – глагол речи, но не является прототипом этого класса), а та, в свою очередь – к экспликации «агенс»;

3) за традиционными экспликациями «агенс», «пациенс» и т.д. оставлены лишь наиболее прототипические случаи фреймовых отношений. В частности, «агенс» - это одушевленный контролирующий ситуацию участник, целенаправленно воздействующий физически на «пациенса» (прототипически - неодушевленный физический предмет, претерпевающий изменение или уничтожение в результате физического воздействия);

4) традиционно выделяемые в лексикологии тематические группы предикатов (например, глаголы речи, восприятия, ментальные предикаты, при-

лагательные оценки и т.п.) имеют в своем ядре единую конфигурацию экспликаций-ролей (ядерный фрейм); таким образом, разметка экспликаций-ролей производится системно по группам предикатов;

5) в языке всегда остается «остаток» - индивидуальные предикаты или синонимические ряды, обозначающие такие конфигурации участников фреймов, которые плохо сводятся к «ядерному» инвентарю ролей. Примеры: субъект в *кашель отпустил, обувь требует ремонта, он проканителился с месяц,* генитивный участник в *практика продаж.* В первую очередь, это касается первого и второго аргумента глагола и первого аргумента имен существительных и прилагательных, интерпретация которых в большей степени зависит от семантики глагола, а не от морфосинтаксического оформления аргумента. Приписывание участникам традиционных ярлыков из узкого списка было бы условностью, присваивание индивидуальных экспликаций – угрожает увеличить инвентарь до бесконечности. Принцип графа позволяет мягко подойти к этой проблеме, отражая условность и градуальность меток.

6) экспликации-роли чувствительны к (не)регулярной полисемии и сдвигу онтологического статуса участников; так, смена роли отражает сдвиг значения и таксономической категории первого актанта в *он занес конверт на почту – ветер занес семена в пустыню.*

7) многие предикаты допускают множественную интерпретацию роли участника, в частности, если в презумптивной и ассертивной части толкования содержатся разные (внутренние) предикаты мнения, речи, социального воздействия и т.п. Соответствующие предикаты известны тем, что принадлежат в этих случаях одновременно к нескольким лексико-семантическим (тематическим) группам глаголов, ср. *мелькать* (движение и восприятие), *заступиться* (говорение и социальные отношения) и др. Мы, однако, стремились выбрать единственную экспликацию, отдавая предпочтение роли участников в ситуации, отраженной в ассертивной части толкования. Об альтернативных и двойных интерпретациях см. ниже.

За основу для составления списка был взят инвентарь семантических ролей, приведенный в Апресян 2010: 370-377, с учетом других работ Московской семантической школы, в частности, словарей ТКС, НОСС, работ Ю.Д.Апресяна, Е.В.Падучевой, Л.Л.Иомдина, Г.И.Кустовой и др., а также англоязычной системы FrameNet. Практическая работа с имеющимися в системе шаблонами конструкций потребовала внесения в список Апресян 2010 ряда изменений.

Во-первых, некоторые роли были объединены в силу незначительности семантических различий между ними, ср. Момент ‘точка или отрезок времени, в которых локализуется какая-то ситуация’ (*начаться в **12 часов**, выезжать в Краков **завтра***) и Дата ‘момент, когда что-то может или должно произойти, или временной отрезок, внутри которого локализуется событие’ (*назначить на **час дня**, перенести на **завтра***).

Во-вторых, оказалось, что ряд ролей из списка Апресян 2010 объединяет достаточно разнородные семантические сущности и соответствующие им разнородные семантические классы глаголов, которые, в свою очередь, было

бы целесообразно различать при разметке для дальнейшего применения системы как в теоретико-семантических исследованиях, так и для решения прикладных задач. Такие роли были разделены нами на несколько экспликаций – например, это коснулось роли Экспериенцера, которой в нашей разметке соответствуют экспликации Субъект восприятия (*видеть, слышать*), Субъект ментального состояния (*думать, понимать*), Субъект психологического состояния (*бояться, любить*), Субъект физиологического ощущения (*болеть, колоть в боку*) и Субъект физиологической реакции (*смеяться, тошнить*), или роли Агенса, которая была сохранена для ядерных агентивных контекстов, но в дополнение к ней в список были включены экспликации Говорящий, Субъект поведения (*лениться, медлить*), Субъект перемещения (последняя экспликация используется для всех, не только агентивных, одноместных глаголов перемещения, коррелируя тем самым с их выделением в особый класс; агентивность глагола в этом случае однозначно устанавливается по одушевленности субъекта).

В-третьих, названия отдельных ролей были изменены в целях придания им большей понятности и самодостаточности (например, вместо ярлыка «Пациенс!» нами использовано наименование «Подвергающаяся воздействию часть пациенса»).

В результате для разметки шаблонов конструкций в текущей версии был использован список порядка 100 экспликаций, классифицированный по принципу семантической близости на несколько групп: «Блок агенса»; «Группа посессивных ролей»; «Блок пациенса»; «Блок экспериенцера»; «Блок инструмента»; «Группа источников и ресурсов»; «Блок обстоятельственных характеристик», далее делящийся еще на несколько подгрупп см. Рис. 1.

При разметке шаблонов конструкций каждой переменной в стандартном случае приписывалась ровно одна экспликация. Возможность приписывания нескольких альтернативных экспликаций (в шаблоне разделяемых знаком «/») допускалась, когда альтернативные экспликации существенно не сдвигают семантики глагола, а выбор между ними часто при этом определяется прагматическими характеристиками ситуации, ср. имеющиеся в одном и том же шаблоне конструкции иллюстрации *Мы рисуем* ***карандашом*** (Инструмент) и *Мы рисуем* ***тушью*** (Средство). В том же случае, если в шаблоне одной конструкции оказывались семантически явно разнородные примеры, производилось разделение этого шаблона на несколько, ср. разнесенные таким образом примеры *Бумага поддается действию огня* (Пациенс; Эффектор), *Саша поддается унынию* (Субъект психологического состояния; Содержание мысли) и *Красота этого уголка не поддается описанию* (Тема; Содержание действия).

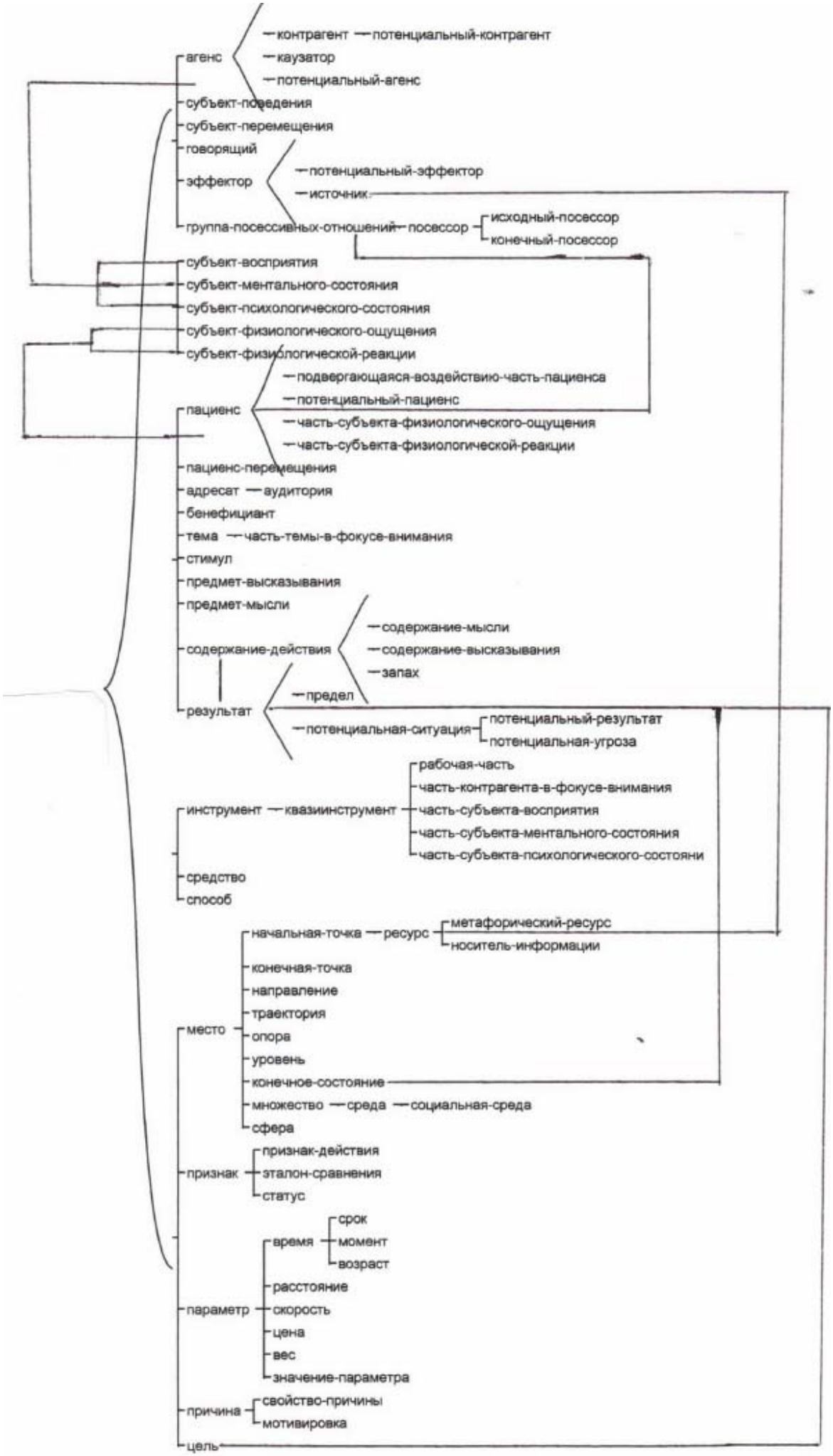

Рис. 1. Иерархия экспликаций (ролей) системы ФреймБанк.

Помимо альтернативных экспликаций, допускалась и возможность приписывания элементу конструкции двойной экспликации (составные части разделены знаком «-») – в том случае, когда этот элемент сочетает в себе семантику каждой из экспликаций. Так, например, в контексте *обрабатывать детали* ***на станке*** речь идет об инструменте совершения действия, но одновременно этот инструмент имеет локативные свойства, поэтому в данном случае использовалась экспликация «инструмент – место». Предложение *Пехотинцы строились* ***клином*** описывает результат (то, что получилось в результате построения) и одновременно способ совершения действия, и в этом и подобных случаях в разметку вводилась двойная экспликация «результат – способ». В ситуации *Мать одела мальчика в тулупчик* пациентивными свойствами обладают два участника – *мальчик* и *тулупчик*, но участник *тулупчик* одновременно имеет и локативную семантику, поэтому участник *мальчик* был размечен как «пациенс», а участник *тулупчик* – как «место – пациенс».

Наряду с указанием экспликаций семантических ролей, были уточнены и внесены в базу также семантические ограничения на заполнение валентностей в шаблонах конструкций базы («лицо», «одушевленный», «неодушевленный», «жидкость», «транспортное средство», «абстрактное имя» и т.д.).

Проведенная работа позволит провести адекватную разметку корпусных примеров, иллюстрирующих лексические конструкции, не только по морфосинтаксическим параметрам, но и по семантическим свойствам участников, что расширит возможности применения этого материала в лингвистических исследованиях.

Итоги работы представлены на международной конференции EURALEX. Подготовлены две статьи в материалах конференции (Lyashevskaya 2012, Ляшевская и др. 2012) и одна статья в сборнике (Митрофанова и др. 2012).

Литература.

Апресян Ю.Д. (отв. ред.). Теоретические проблемы русского синтаксиса: Взаимодействие грамматики и словаря. – М.: ЯСК, 2010. С. 370 – 377.

Lyashevskaya Olga. Dictionary of Valencies Meets Corpus Annotation: A Case of Russian FrameBank // Proceedings of EURALEX 2012, Oslo, Norway.

Ляшевская, О. Н., О. А. Митрофанова, М. А. Грачкова, А. С. Шиморина, А. С. Шурыгина, С. В. Романов. К построению инвентаря русских именных конструкций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая - 3 июня 2012г.). Вып. 11 (18).- М.: Изд-во РГГУ, 2012. С. 370–382.

Митрофанова, О. А., О. Н. Ляшевская, М. А. Грачкова, А. С. Шиморина, А. С. Шурыгина, С. В. Романов. Автоматическое разрешение лексико-семантической неоднозначности и выделение конструкций (на материале Национального корпуса русского языка) // Лексикология. Лексикография. Корпусная лингвистика. ИЛИ РАН. СПб (в печати).

6. Общее число опубликованных в 2012 г. по проекту работ
6.1. количество монографий 0
6.2. количество сборников статей 0
6.3. количество статей 3 (две статьи в материалах конференции, одна статья в сборнике).

7. Список опубликованных монографий и сборников статей, с полным указанием выходных данных, объема /в п.л. и количество стр./, а также их краткие аннотации (до 0,5 стр.)

8. Список опубликованных по проекту статей (объемом не менее 1 п.л.)

9. Список книг, сданных в печать или поданных на издательские гранты (указать ожидаемое время издания, объем в а.л.)

10. Экспедиции, проведенные в рамках проекта (регион, руководитель, сроки, тематика исследований, полученные результаты, их значимость – до 2 стр.)

11. Конференции, организованные в рамках проекта (название, место и сроки проведения, обсуждаемые проблемы, результаты)

12. Важнейшие научные результаты работы по проекту (ок. 0,5 стр. для публикации на сайте Программы)

Создана пилотная онлайн-версия системы «ФреймБанк» в режиме «словарь» и «размеченные примеры». Система пополнена представительными выборками примеров из НКРЯ (по 100 примеров) для 500 глаголов и предикатных имен. Примеры размечены по следующей схеме: каждый пример соотносится с соответствующей конструкцией в словаре, аннотируется выражение/невыражение участника, синтаксические особенности кодирования (ранг, связь с предикатом через контроль, изменение ранга в связи с пассивной, аттрибутивной причастной и др. типами конструкций и т.п.), особенности предложно-падежного и др. морфосинтаксических способов оформления участников; также аннотируются сирконстантные участники ситуации и модальные элементы. Отредактирована разметка корпусных данных для 1000 ранее размеченных лексических единиц.

Составлена иерархия глобальных и периферийных семантических ролей, учитывающая данные по 1500 лексическим единицам русского языка. Иерархия будет дорабатываться с учетом включения новых данных в систему. Иерархия представляет собой связанный граф, включающий ядерную часть (такие «классические» роли как «Агенс», «Экспериенцер», «Адресат», «Пациенс» и др.), а также более частные типы ролей (экспликации), встречающиеся в конкретных тематических группах глаголов и даже на уровне индвидуальных семантических рядов и лексем. Так, например, глобальной роли Экспериенцера соответствуют более частные роли Субъект восприятия (ср. *видеть, слышать*),

Субъект ментального состояния (ср. *думать, понимать*), Субъект психологического состояния (ср. *бояться, любить*), Субъект физиологического ощущения и т.п. За ролью Пациенса на частном уровне иерархии оставлены прототипические случаи его реализации в ситуациях физического уничтожения и разрушения.

Итоги работы представлены на международной конференции EURALEX.

13. Наиболее значимый научный результат проекта (5-6 строк для сводного отчета в Президиум РАН)

Создана пилотная онлайн-версия системы «ФреймБанк». В систему включены представительные выборки примеров из НКРЯ для 1500 глаголов и предикатных имен, в каждом примере указан тип фреймовой конструкции ключевого предиката в словаре системы, размечены семантические роли и морфосинтаксическое оформление участников фрейма. Составлена иерархия глобальных и периферийных семантических ролей.

14. Краткий финансовый отчет за 2012 г. (основные статьи расходов по проекту, сумма)

15. Запрашиваемый на 2013 г. объем финансирования, с кратким обоснованием расходов

16. Краткое обоснование научных работ на 2013 г., ожидаемые результаты – до 1 стр. и заполненная форма 2 (см. ниже).

1. В 2013 году предполагается провести эксперименты по а) автоматическому составлению сбалансированных выборок; выбору параметров случайной выдачи и сбалансированной выдачи;
б) автоматическому распознаванию участников и семантических ролей (semantic role labeling) по выборкам с приписанными лексическими конструкциями;
в) автоматическому распознаванию типа конструкции в результате обучения с учителем – на выборках ранее размеченных конструкций для частотных предикатов.
Будут проведены эксперименты по автоматическому заполнению части шаблона реализации конструкции с учетом корпусных данных.
2. Будет проведена экспертиза иерархии семантических ролей (экспликаций) а) в ходе консультаций с экспертами в области; б) с учетом статистической модели, построенной на корпусных данных, с последующим ручным редактированием выпадающих из модели случаев.
3. Будет составлен словарь моделей управления русских предикатов (глаголов, имен) с разметкой а) базового морфосинтаксического шаблона, б) синтаксического ранга участников, в) экспликации участников; г) ограничений на заполнение элементов конструкции (пилотная версия, основанная на ранее размеченных данных). В словарь будет соддержать ссылки на внешние ресурсы: словарь системы ЭТАП-2 (частично), словарь МАС (частично),

FrameBank (экспериментальная версия на части данных), другие онлайн-ресурсы.
4. Модель многозначности 100 глаголов и их гипо-гиперонимических отношений будет интерпретирована с помощью сходства семантики фреймов и мотивации морфо-синтаксического оформления лексических конструкций.
5. Разметка реализаций конструкций в текстовых примерах будет углублена за счет разметки лексико-семантического класса слов, замещающих позицию того или иного актанта. В дальнейшем планируется собрать данные для разметки соответствующего компонента в словаре валентностей, а также данные о нестандартном заполнении валентностей.

Ожидаемые результаты:
- словарь моделей управления русских предикатов (с учетом корпусных данных) – около 2000 лексических единиц;
- составление графа фреймов для 100 лексических единиц: наследование и другие типы связей между фреймами, стоящими за разными предикатами и разными значениями предикатов;
- разметка в системе ФреймБанк расширенных выборок для частотной русской предикатной лексики (5000 примеров);
- редактирование ФреймБанк-разметки корпусных данных и инвентаря семантических ролей (на материале 1500 ранее размеченных лексических единиц)

Подпись руководителя проекта Е.В. Падучева

Форма 2
Планируемое содержание работ на 2013 г.

№	Название проекта	Организация-исполнитель и учреждения-соисполнтели	Руководитель проекта (+ кол-во исполнителей)	Запрашиваемый объем финансирования на 2013 г. (тыс. руб.)	Ожидаемые в 2013 г. результаты
	ФреймБанк: разметка семантических ролей и морфосинтаксического оформления участников фреймов (на базе НКРЯ)	ИРЯ РАН	Падучева Е.В. (+ 11)		- словарь моделей управления русских предикатов (с учетом корпусных данных) – около 20000 входов; - составление графа фреймов для 100 лексических единиц: наследование и другие типы связей между фреймами, стоящими за разными предикатами и разными значениями предикатов; - разметка в системе ФреймБанк расширенных выборок для частотной русской предикатной лексики (5000 примеров); - редактирование ФреймБанк-разметки корпусных данных и инвентаря семантических ролей (на материале 1500 ранее размеченных лексических единиц)