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
КОРПУСНАЯ ЛИНГВИСТИКА. Направление 1. Создание и развитие корпусных ресурсов по современному русскому языку.</td></tr>
<tr><td colspan="2">Название проекта
Развитие глубоко аннотированного корпуса текстов «СинТагРус» с созданием подкорпуса эллиптических конструкций русского языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Иомдин Леонид Лейбович, кандидат филологических наук</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города), почтовый адрес руководителя проекта</td><td>e-mail: iomdin@iitp.ru; тел. (495) 699-49-27; факс (495)650-05-79, 127994, Москва, ГСП-4, Большой Каретный пер. 19, стр. 1, ИППИ РАН</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования

Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН)</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Кулешов Александр Петрович, академик РАН</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Лобачёва Римма Ивановна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
127994, Москва, ГСП-4, Б.Каретный пер., д. 19, стр. 1, ИППИ РАН; тел. (495) 650-42-25; факс (495)650-05-79;
e-mail: director@iitp.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 rowspan="5">Исполнители (ФИО, уч. ст.)</td><td>Лазурский Александр Вадимович</td></tr>
<tr><td>Митюшин Леонид Григорьевич, к.ф.-м.н.</td></tr>
<tr><td>Подлесская Ольга Юрьевна, к.филол.н.</td></tr>
<tr><td>Фролова Татьяна Ильинична</td></tr>
<tr><td>Цинман Леонид Львович, к.ф.-м.н.</td></tr>
<tr><td>Дата сдачи отчета
20 ноября 2012 г.</td><td>Подпись руководителя проекта:</td></tr>
</table>

Форма 1
УТВЕРЖДАЕМ

Координатор Программы акад. РАН Вяч. Вс. Иванов

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

«___»_____ 2012 г.

1. Название направления
Направление 1. Создание и развитие корпусных ресурсов по современному русскому языку.
2. Название проекта
Развитие глубоко аннотированного корпуса текстов «СинТагРус» с созданием подкорпуса эллиптических конструкций русского языка.
3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)
Иомдин Леонид Лейбович, кандидат филологических наук, ведущий научный сотрудник, ИППИ им. А.А. Харкевича РАН
4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)
Лазурский Александр Вадимович, старший научный сотрудник, ИППИ им. А.А. Харкевича РАН;
Митюшин Леонид Григорьевич, кандидат физико-математических наук, старший научный сотрудник, ИППИ им. А.А. Харкевича РАН;
Подлесская Ольга Юрьевна, кандидат филологических наук, научный сотрудник, ИППИ им. А.А. Харкевича РАН;
Фролова Татьяна Ильинична, младший научный сотрудник, ИППИ им. А.А. Харкевича РАН;
Цинман Леонид Львович, кандидат физико-математических наук, ведущий научный сотрудник, ИППИ им. А.А. Харкевича РАН.
5. Содержание фактически проделанной работы, полученные результаты, их новизна, научная и практическая значимость, актуальность, соответствие мировому научному уровню (до 5 стр.)
Важнейшим содержанием работ, проведенных в 2012 году, является дальнейшее развитие корпуса СинТагРус, содержащего русские тексты, снабженные полной морфосинтаксической разметкой. Общий объем корпуса составил около 53 400 предложений (свыше 770 000 словоупотреблений), из которых 4020 предложений (около 65 800 словоупотреблений) были обработаны и добавлены в корпус в отчетном году. Большой объем корпуса и детальность содержащегося в нем материала делают его весьма ценным и уникальным ресурсом, который может быть успешно использован в широком классе исследовательских и практических задач компьютерной лингвистики.

Включенные в корпус в 2012 г. 28 новых текстов относятся к научно-популярному, публицистическому, биографическому и новостному жанрам. Выбор текстов этих жанров позволяет отразить в синтаксически аннотированном корпусе современное

состояние русского литературного языка и в то же время ограничить появление в корпусе материала, синтаксическая кодификация которого затруднительна (разговорная речь, поэзия, диалектные тексты, техническая документация). Как следствие, материал корпуса становится синтаксически более однородным, что приводит к повышению качества синтаксической разметки и достижению большего единообразия принимаемых лингвистических решений.

При построении корпуса СинТагРус принята следующая процедура. Вначале непрепарированный текст обрабатывается программой сегментации текста, которая автоматически разбивает его на отдельные предложения. После этого каждое предложение пропускается через парсер многофункционального лингвистического процессора ЭТАП-3, в результате чего вырабатывается его морфосинтаксическая структура (дерево зависимостей). Парсер выполняет также лексико-семантическую и лексико-функциональную разметку предложения. Лексико-семантическая разметка заключается в том, что многозначным словам приписывается одно из значений, имеющихся в комбинаторном словаре процессора ЭТАП-3. В ходе лексико-функциональной разметки обнаруживаются и отмечаются словосочетания, допускающие интерпретацию в терминах лексических функций; при этом также используется информация, содержащаяся в комбинаторном словаре. Затем структуры, построенные системой ЭТАП-3, проверяются экспертами-лингвистами.

Фактически почти все затраты труда при разработке корпуса приходятся на последний этап работы, выполняемый специально подготовленными аннотаторами. Хотя лингвистический процессор ЭТАП-3 насыщен весьма богатой и разнообразной лингвистической информацией, он не способен строить морфосинтаксические структуры со стопроцентной надежностью. Аннотаторы проверяют все элементы структур, созданных парсером процессора, и вносят необходимые коррективы с помощью специального программного комплекса "Редактор структур" (основная составная часть программной среды "Рабочее место аннотатора").

Использование процессора ЭТАП-3 для разработки корпуса создает обратную связь, важную для самого процессора, поскольку при выполнении этой работы база русских грамматических правил и русский словарь процессора постоянно пополняются и совершенствуются. Отметим, что процессор рассчитан на обработку русских и английских текстов без каких-либо тематических и лексико-грамматических ограничений. Он также включает менее детально разработанные компоненты для работы с текстами на французском, немецком, испанском, арабском и корейском языках. Русский морфологический словарь в настоящее время содержит около 130 000 входов, а комбинаторный словарь – более 100 000 входов.

Одной из наиболее трудных задач, возникающих при автоматическом синтаксическом анализе, является обработка эллиптических конструкций. Предложения с эллипсисом составляют 2–3 процента от общего числа предложений корпуса СинТагРус. Они представлены в корпусе деревьями зависимостей, в которых эллиптированным словам соответствуют особые узлы дерева, имеющие пустой текстовый элемент. В настоящее время парсер лингвистического процессора ЭТАП-3 не рассчитан на обработку эллиптических предложений, так как не располагает надежными средствами восстановления

эллиптированных элементов. В рамках данного проекта предполагается уделить эллиптическим конструкциям в корпусе особое внимание, поскольку более полное понимание их особенностей может стать основой для создания эффективных алгоритмов анализа таких конструкций.

Построение аннотированных корпусов текстов является весьма актуальной и быстро развивающейся областью современной компьютерной лингвистики. Это обусловлено тем, что при создании систем автоматической обработки текстов все более широко применяются методы машинного обучения, и размеченные корпусы представляют собой именно те массивы данных, на которых проводится обучение. В настоящее время существует не менее 70 корпусов для различных языков, в которых аннотация достигает синтаксического уровня (в том числе не менее 15 для английского языка).

Новизна результатов, связанных с построением корпуса СинТагРус, определяется тем, что этот корпус является единственным в мире большим аннотированным корпусом текстов на русском языке с аннотацией на морфосинтаксическом уровне. Другой важной особенностью данного корпуса является представление синтаксических структур предложений в виде деревьев зависимостей. При этом различается около 70 типов синтаксических связей между словами, что обеспечивает более полное и лингвистически содержательное представление синтаксической структуры, чем в других существующих корпусах с синтаксической разметкой.

Как и в предыдущие годы, в 2012 г. корпус СинТагРус регулярно применялся для регрессионного тестирования парсера процессора ЭТАП-3. Идея регрессионного тестирования состоит в том, что некоторая часть корпуса (порядка 15 тысяч предложений) принимается за эталон. Эти предложения пропускаются в автоматическом режиме через парсер системы, и фиксируются все расхождения между структурами, полученными автоматически, и структурами корпуса. Через определенные промежутки времени автоматический анализ данного набора предложений повторяется, и его результаты снова сравниваются с эталонным анализом. Динамика расхождений наглядно показывает, как меняется работа парсера в результате модификаций, внесенных в лингвистическое и программное обеспечение системы, и позволяет немедленно исправлять ошибочные или недостаточно надежные решения.

В 2012 г. получил дальнейшее развитие подход, связанный с использованием статистических данных корпуса СинТагРус в парсере лингвистического процессора ЭТАП-3. На основании данных корпуса вырабатываются оценки вероятностей альтернативных вариантов для элементов морфосинтаксической структуры, а именно оценки вероятностей различных лексико-морфологических интерпретаций одного и того же слова предложения и оценки вероятностей различных синтаксических связей, входящих в один и тот же узел дерева зависимостей. На основе этих данных парсер ищет дерево зависимостей, удовлетворяющее всем требованиям грамматических правил системы и имеющее максимальную результирующую оценку вероятности. Эксперименты показали, что этот подход заметно повышает качество синтаксических структур, получаемых парсером.

6. Общее число опубликованных в 2012 г. по проекту работ

6.1. количество монографий

6.2. количество сборников статей
6.3. количество статей - 2

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.)

1. Leonid Iomdin, Vadim Petrochenkov, Victor Sizov, Leonid Tsinman. ETAP parser: state of the art. (Л.Л.Иомдин, В.В.Петроченков, В.Г. Сизов, Л.Л.Цинман Синтаксический анализатор системы ЭТАП: современное состояние). // Dialog 2012. Computational Linguistics and Intellectual Technologies. (International Conference. Moscow, RGGU Publishers, Issue 11(18). P. 830-843 (на англ. языке)
2. Leonid Iomdin. Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact. Jazykovedné štúdie, Bratislava (на англ. языке, в печати).

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)

Синтаксически размеченный корпус СинТагРус является уникальной научной разработкой, основное содержание которой заключается в применении модели естественного языка "Смысл ⇔ Текст" для интерпретации реальных русских текстов во всем их смысловом и структурном многообразии. В 2012 г. корпус был пополнен 28 текстами общим объемом 4020 предложений. Для этих текстов были построены морфосинтаксические структуры предложений и маркированы лексические функции. С учетом новой порции, общий объем корпуса составил около 53 400 предложений (более 770 000 тысяч словоупотреблений).

Статистические данные корпуса применяются в парсере лингвистического процессора ЭТАП-3. Парсер вычисляет оценки вероятностей альтернативных элементов морфосинтаксической структуры и ищет дерево зависимостей, удовлетворяющее всем требованиям грамматических правил системы и имеющее максимальную результирующую оценку вероятности.

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

В 2012 г. корпус был пополнен 28 текстами общим объемом 4020 предложений. Для этих текстов были построены морфосинтаксические структуры предложений и маркированы лексические функции. С учетом новой порции, общий объем корпуса составил около 53 400 предложений (более 770 000 тысяч словоупотреблений).

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).

В 2013 году будет производиться пополнение и развитие корпуса СинТагРус. Общий объем корпуса составит не менее 56500 предложений (не менее 830000 слов).

Для достижения этой цели будет совершенствоваться автоматический синтаксический анализатор, с помощью которого производится первичная разметка СинТагРуса (это усовершенствование, в частности, будет касаться разработки средств частичной автоматизации синтаксического анализа некоторых типов эллиптических конструкций, что может потребовать изменений не только в используемых лингвистических правилах, но и в алгоритме анализа и в программном комплексе в целом).

Автоматический комбинаторный словарь русского языка, используемый синтаксическим анализатором, будет систематически пополняться, в первую очередь, за счет значений лексических функций ключевых слов, а также за счет разукрупнения значений некоторых многозначных слов.

Предполагается написание двух-трех научных статей и выступление с докладами, на конференциях по корпусной и компьютерной лингвистике.

Руководитель проекта
к.филол.н. Иомдин Л.Л.

Форма 2
Планируемое содержание работ на 2013 г.

№	Название проекта	Организация-исполнитель и учреждения-соисполнители	Руководитель проекта (+ кол-во исполнителей)	Запрашиваемый объем финансирования на 2013 г. (тыс. руб.)	Ожидаемые в 2013 г. результаты
36П	Развитие глубоко аннотированного корпуса текстов «СинТагРус» с созданием подкорпуса эллиптических конструкций русского языка	ИППИ РАН	Иомдин Л.Л. + 5		Увеличение объема корпуса до 56,5 тыс. предложений. Расширение подкорпуса эллиптических конструкций до объема в 2300 предложений.