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 Создание и развитие корпусных ресурсов по древнерусскому языку</td></tr>
<tr><td colspan="2">Название проекта Подкорпус древнерусских текстов XI-XIII вв.</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Пичхадзе Анна Абрамовна, кфн</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>rusyaz@yandex.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования
Институт русского языка им. В. В. Виноградова РАН (ИРЯ РАН)</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
+7(495)695-28-07</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>Архангельский Тимофей Александрович</td></tr>
<tr><td>Макеева Ирина Ивановна</td></tr>
<tr><td>Мушинская Мария Савельевна</td></tr>
<tr><td>Баранкова Галина Серафимовна</td></tr>
<tr><td>Петрухин Павел Владимирович</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 Создание и развитие корпусных ресурсов по древнерусскому языку

2. Название проекта Подкорпус древнерусских текстов XI-XIII вв.

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы) Пичхадзе Анна Абрамовна, кфн, внс ИРЯ РАН

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

Архангельский Тимофей Александрович, преподаватель отделения лингвистики факультета филологии НИУ ВШЭ

Макеева Ирина Ивановна, кфн, ИРЯ РАН, снс

Мушинская Мария Савельевна, ИРЯ РАН, нс

Баранкова Галина Серафимовна, кфн, ИРЯ РАН, снс

Петрухин Павел Владимирович, кфн, ИРЯ РАН, снс

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

Древнерусский подкорпус в составе НКРЯ в 2012 г. пополнялся за счет оригинальных древнерусских сочинений, выполненных на Руси переводов и церковнославянских памятников южнославянского происхождения. Пополнение составило ок. 170000 словоупотреблений. Из оригинальных древнерусских памятников в подкорпус была включена единственная из древнерусских летописей, отсутствовавшая в нём раньше, - Повесть временных лет (ок. 54500 словоупотреблений). Эта древнейшая восточнославянская летопись представлена в корпусе не по Лаврентьевскому списку 1377 г., по которому она издана со словоуказателем, а по Ипатьевскому списку ок. 1425 г., указатель к которому до сего времени отсутствовал. Кроме того, в подкорпус были включены три нарративные произведения древнерусского писателя XII в. Кирилла Туровского. Разметке текстов предшествовала текстологическая работа с рукописными источниками с целью установления наиболее аутентичных источников. Надежно атрибутируются Кириллу следующие сочинения, подписанные в древнерусских рукописных памятниках его именем: «Сказание о черноризском чине», «Притча о душе и теле», «Повесть о беспечном царе и мудром советнике», 8 «Слов» и молитвы. В текущем году размечены и включены в подкорпус первые три памятника (общий объем ок. 8500 словоупотреблений): аскетическое произведение «Сказание о черноризском чине», возможно, предназначенное для иноков туровского Борисоглебского монастыря, – база данных для подкорпуса подготовлена по древнейшему сохра-

нившемуся списку в составе Новгородской кормчей конца XIII в., до сих пор не опубликованному, содержащему так называемую «редакцию кормчих»; «Притча о душе и теле» – база данных подготовлена по списку ГИМ, Чуд. 20, XIV в.; «Повесть о беспечном царе и его мудром советнике» – база данных выполнена по ранее не публиковавшемуся списку ГИМ, Синод. собр. № 935, XVI в., в котором, несмотря на позднюю дату списка, представлена редакция памятника, в значительной степени сохранившая полный аутентичный авторский текст и архаичную лексику.

Из переведенных в Древней Руси памятников в текущем году в подкорпус была включена «История Иудейской войны» Иосифа Флавия (ок. 84000 словоупотреблений). Специфика работы над этим текстом была обусловлена тем, что в отличие от остальных текстов он не размечался в виде базы данных. Грамматическая разметка текста "Истории Иудейской войны" была произведена автоматически на основании электронного варианта словоуказателя к этому тексту в формате Word, содержащего все грамматические пометы и греческие соответствия. Была специально разработана программа, позволившая преобразовать этот словоуказатель в базу данных в формате Morphy, которая и была включена в подкорпус. Часть возникшей при этом омонимии была снята вручную. Оставшиеся словоформы с несколькими омонимичными разборами в основном представляют собой случаи, когда два омонима содержатся на одной и той же строке текста; такие словоформы составляют всего 0,4% от общего объёма текста. Этот беспрецедентный опыт открывает возможности превращения хорошо структурированных указателей традиционных печатных изданий в базы данных с их последующим использованием в Корпусе.

Из церковнославянских памятников южнославянского происхождения в подкорпус включен Изборник 1076 г. (ок. 20000 словоупотреблений) – сборник нравственно-назидательных текстов различных жанров. Основу сборника составляют переводы с греческого (библейская книга Премудрости Иисуса сына Сирахова, произведения византийских писателей), несколько текстов («Слово некоего отца к сыну», «Стословец», «Слово о чтении книг»), по-видимому, являются оригинальными славянскими сочинениями. Рукопись является древнерусской по письму, но переписанные в ней тексты созданы в Болгарии в X в. Изборник 1076 г. является самым старшим памятников из всех, входящих в подкорпус.

Тексты нескольких памятников (Повести временных лет, сочинений Кирилла Туровского, Изборника 1076 г.), которые вместе с грамматической разметкой хранились в базах устаревших форматов, отличающихся друг от друга и основанных на MS Access, были переведены в формат YAML, используемый в настоящий момент для хранения и разметки текстов. Перевод текстов в данный формат решает многие технические проблемы (особенно связанные с разметкой разрывных словоформ типа глаголов с частицей *-ся* и аналитических форм), существенно уменьшает вероятность появления ошибок в базе и необходим для размещения их в Национальном корпусе русского языка. Перевод текстов осуществлялся в полуавтоматическом режиме: система, созданная для этих целей, извлекает информацию из базы в формате MS Access, параллельно сообщая об ошибках в базе (например, различный внешний вид словоформы в строке памятника и в таблице с грамматическими разборами), которые исправляются вручную.

К модулю, осуществляющему конвертирование размеченных текстов в формат XML, используемый в Национальном корпусе русского языка, был добавлен ряд новых функций (например, запись информации о значении слова, позволяющей развести омонимы при поиске). Все перечисленные тексты были преобразованы в формат XML с помощью этого модуля и загружены в Национальный корпус русского языка.

В текущем году был несколько усовершенствован формат грамматической разметки – в частности, оптимизирован разбор аналитических форм типа перфекта и добавлены некоторые пометы, специфические для языка восточнославянских памятников. После разметки тексты были снабжены таблицами с метаинформацией и переданы для вывески в Национальный корпус русского языка в формате, позволяющем осуществлять поиск по этим текстам с использованием средств Корпуса.

6. Общее число опубликованных в 2012 г. по проекту работ
6.1. количество монографий
6.2. количество сборников статей
6.3. количество статей

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)

В 2012 г. тексты памятников, хранящиеся в виде баз данных MS Access, были переведены в новый формат и размечены (или доразмечены) с использованием программы Morphy: каждой словоформе были приписаны грамматические признаки, дополнительные пометы, а для переводных текстов — переводной эквивалент. В корпус включены древнейшая восточнославянская летопись Повесть временных лет, три нарративных сочинения Кирилла Туровского, древнерусский переводы с греческого – «История Иудейской войны» Иосифа Флавия, а также церковнославянский переводной памятник южнославянского происхождения Изборник 1076 г. После разметки тексты были снабжены таблицами с метаинформацией и переданы в Национальный корпус русского языка в формате, позволяющем осуществлять поиск по этим текстам с использованием средств Корпуса. Поиск по текстам (в том числе с использованием упрощённой орфографии) доступен в Историческом подкорпусе.

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

С помощью среды для ручной и автоматической (на основе уже обработанных файлов) обработки древнерусских текстов Morphy был размечен ряд оригинальных и переводных памятников XI-XIII вв. Создана программа преобразования грамматических словоуказателей в формате Word в базы данных в среде Morphy. Размеченные тексты переданы в Национальный корпус русского языка и доступны для поиска в Историческом подкорпусе. Корпус отражает весь лексикон размеченных текстов, в том числе данные по ономастике и топонимии, не отраженные историческими словарями.

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 г. планируется добавить в подкорпус материалы древнерусских переводов «Жития Андрея Юродивого» (ок. 45000 словоупотреблений) и «Повести об Акире Премудром» (более 8000 словоупотреблений), Галицкого евангелия 1144 г. (ок. 40000 словоупотреблений) и некоторых текстов Кирилла Туровского.

Подпись руководителя проекта Пичхадзе А. А.

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
	Подкорпус древнерусских текстов XI-XIII вв.	ИРЯ РАН	к.ф.н. А.А. Пичхадзе + 5 исполнителей		2013 г. — Добавление в подкорпус «Жития Андрея Юродивого» (ок. 45000 словоупотреблений), «Повести об Акире» (более 8000 словоупотреблений), Галицкого евангелия 1144 г. (ок. 40000 словоупотреблений) и некоторых текстов Кирилла Туровского.