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 истории русского языка</td></tr>
<tr><td colspan="2">Название проекта

Электронный корпус древненовгородских письменных источников: летописи, деловые и юридические памятники</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Гиппиус Алексей Алексеевич, д.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>agippius@mail.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования

Учреждение Российской Академии наук «Институт славяноведения РАН»</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Никифоров Константин Владимирович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Боченова Наталья Владимировна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
Москва, 119336, Ленинский пр. 32а,
(495) 938 17 80
(499) 938 00 96
inslav@inslav.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>Мишина Екатерина Андреевна, к. ф.н.</td></tr>
<tr><td>Минлос Филипп Робертович, к.ф.н.</td></tr>
<tr><td>Санников Андрей Владимирович, к.ф.н.</td></tr>
<tr><td>Трефилова Ольга Владимировна</td></tr>
<tr><td>Архангельский Тимофей Александрович</td></tr>
<tr><td>Дата сдачи отчета</td><td>20.11.2012</td></tr>
<tr><td></td><td></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

Направление 2. Создание и развитие корпусных ресурсов по древнерусскому языку.

2. Название проекта

Электронный корпус древненовгородских письменных источников: берестяные грамоты и Новгородская первая летопись.

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)

Гиппиус Алексей Алексеевич, д-р филол. наук, вед. научн. сотруд., Институт славяноведения РАН

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

Архангельский Тимофей Александрович, преподаватель Национально го исследовательского университета «Высшая школа экономики»

Минлос Филипп Робертович, к.ф.н, науч. сотр. Института славяноведения РАН

Мишина Екатерина Андреевна, к.ф.н, науч. сотр. Института русского языка **РАН**

Санников Андрей Владимирович, к.ф.н, мл. науч. сотр. Института русского языка РАН

Трефилова Ольга Владимировна, н.с. Института славяноведения РАН

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

В 2012 г. была завершена работа по представлению в НКРЯ текстов **новгородских берестяных грамот** и **старшего извода Новгородской первой летописи (Синодального списка НПЛ)**. Работа с грамотами заключалась в основном в решении ряда технических проблем, выявленных в ходе тестирования предварительной версии, размещенной в НКРЯ в конце 2011 г; работа с летописью – в ручной морфологической разметке остававшейся неразмеченной части текста, и редактировании данных. Морфологическая разметка текстов грамот и летописи производилась вручную. Для обработки текстов грамот и Новгородской первой летописи использовалась среда Morphy, в которую при необходимости вносились изменения и доработки. К имеющейся базе с грамотами были добавлены новые тексты из раскопок 2010 г. (№ 974-1015). Эти грамоты были переведены из текстового формата в формат YAML, используемый для хранения текстов с грамматической разметкой в среде Morphy, после чего были размечены. Над текстами грамот и Новгородской первой летописи было произведено несколько автоматических преобразований, связанных с унификацией грамматической разметки (например, падежные пометы, ранее использовавшиеся для указания информации о предложном управлении, были заменены на специальные пометы об управлении). Информация, частично отсутствовавшая в базе с грамотами -- комментарии и некоторые грамматические пометы, -- была автоматически перенесена в неё из старой базы формата MS Access.

Для работы с текстами грамот использовался шрифт NovgorodATE, в который было включено множество специальных символов, встречающихся в текстах грамот. Однако работа с ним вызывала сложности, поскольку коды используемых в нём символов не соответствовали стандарту Unicode, из-за чего приходилось использовать специальные способы ввода информации с клавиатуры и конвертировать базу при выгрузке в корпус. Чтобы решить эти проблемы, была создана новая версия шрифта, в которой все символы, описанные в стандарте Unicode, расположены на правильных местах. База с грамотами с помощью специального скрипта была переработана для использования с этим шрифтом. Кроме того, в модуль выгрузки текстов в корпус была встроена замена некоторых нестандартных символов на упрощённые варианты для облегчения поиска; например, титла разного вида были заменены на одно стандартное (U+0483).

База с грамотами и Новгородской первой летописью были конвертированы в формат XML, используемый в Национальном корпусе русского языка, и загружены в корпус.

Параллельная велась работа по подготовке к размещению в корпусе других текстов новгородского происхождения. Были выверены по изданиям электронные тексты **Новгородской первой летописи младшего извода** и «**Грамот Великого Новгорода и Пскова**» (ГВНП). Планировавшаяся на 2012 г. подготовка новой, лингвистически более точной транскрипции ГВНП, была отложена до осуществления автоматизированной морфологической разметки существующего электронного текста.

Был также подготовлен по рукописи (ГИМ., Син. 132, Новгородская Кормчая 1282 г.) текст важнейшего памятника новгородской словесности XII в. - **«Вопрошания Кирика»**. Ввиду особой значимости этого памятника, его морфологическая разметка будет производиться вручную.

В рамках проекта велась также работа по разработке морфологического анализатора (парсера), который может использоваться для автоматического анализа текстов 11-17 века. Для этого необходимо создать:

1) компьютерный грамматический словарь, содержащий основную лексику указанного периода с указанием информации о грамматике и словоизменении;
2) формальное описание словоизменения, охватывающее, как минимум, наиболее частотные и регулярные типы.

В качестве первого шага на пути создания грамматического словаря был обработан словник Словаря 11-17 века. Словник, охватывавший материал 1-25 томов издания, был дополнен на основе 26-30 томов, а для оставшейся части алфавита (Т-Я) – на основе «Материалов для словаря древнерусского языка И. И. Срезневского». Полученный таким образом словник был обработан следующим образом: для слов, совпадающих с современным языком, была приписана грамматическая информация из современных словарей, а для слов, отсутствующих в современном языке, были построены гипотезы по аналогии с существующими словами (по характерному концу слова). В данный момент проводится проверка автоматически приписанной информации с целью уточнения и исправления ошибок.

Прежде всего, проверяется часть речи и основные грамматические признаки. На следующем этапе лексемам будут приписываться словоизменительные типы. Параллельно будет создаваться формальное описание словоизменения в виде таблиц парадигм.

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)

Подготовлен и размещен в Интернете текст старшего извода (Синодального списка) Новгородской первой летописи XIII-XIV вв. с полной морфологической разметкой (32 000 словоформ). Создан соответствующий стандарту UNICODE шрифт для представления в Интернете текстов новгородских берестяных грамот. Подготовлен по рукописи Новгородской Кормчей 1282 г. для размещения в Интернете текст «Вопрошания Кирика» XII в.; произведена обработка дополненного словника Словаря русского языка XI-XVII в. для создания на его основе морфологического анализатора древнерусских и старорусских текстов.

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

Подготовлен и размещен в Интернете текст старшего извода (Синодального списка) Новгородской первой летописи XIII-XIV вв. с полной морфологической разметкой (32 000 словоформ).

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

14.1. Заработная плата с начислениями (ФИО, 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. в рамках выполнения проекта планируются следующие работы:

а) пополнение корпуса берестяных грамот текстами грамот № 1016-1051 из раскопок 2011-2012 г.; внесения исправлений в электронный корпус грамот на основании подготовленного XII тома серии «Новгородские грамоты на бересте»;

б) морфологическая разметка текста Новгородской первой летописи младшего извода (90 000 словоупотреблений, полуавтоматизированная);

в) морфологическая разметка «Грамот Великого Новгорода и Пскова» (105 000 словоупотреблений, полуавтоматизированная);

г) морфологическая разметка «Вопрошания Кирика» по списку Новгородской Кормчей 1282 г. (вручную).

Подпись руководителя проекта А. А. Гиппиус

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
	Электронный корпус древненовгородских письменных источников: летописи, деловые и юридические памятники	ФГБУН Институт славяноведения РАН	А.А.Гиппиус 5 исполнителей		Представление в Интернете текста Новгородской первой летописи младшего извода с морфологической разметкой (90 000 словоформ) Представление в Интернете «Грамот великого Новгорода и Пскова с морфологической разметкой (100000 словоформ) Представление в интернете текста «Вопрошания Кирика» с морфологической разметкой (6000)