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 Направление 2. Создание и развитие корпусных ресурсов по древнерусскому языку.</td></tr>
<tr><td colspan="2">Название проекта Развитие корпуса церковнославянских текстов</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Плетнева Александра Андреевна, к.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>krav62@mail.ru, edobrush@gmail.com</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 Москва, 121019, ул. Волхонка, 18/2, (495) 695-26-60 Факс: (495) 695-26-03 E-mail: irlras@mail.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="4">Исполнители (ФИО, уч. ст.)</td><td>Кравецкий Александр Геннадьевич, к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Добрушина Екатерина Роландовна, к.ф.н., ПСТГУ</td></tr>
<tr><td>Поляков Алексей Евгеньевич, НПБ им. К.Д. Ушинского</td></tr>
<tr><td>Иванова-Алленова Татьяна Юрьевна, к.ф.н., ПСТГУ</td></tr>
<tr><td>Дата сдачи отчета
30.11.2012</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 **Направление 2. Создание и развитие корпусных ресурсов по древнерусскому языку.**

2. Название проекта **Развитие корпуса церковнославянских текстов**

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

Плетнева Александра Андреевна, к.ф.н., ИРЯ им. Виноградова, снс

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

Кравецкий Александр Геннадьевич, к.ф.н., ИРЯ РАН, снс

Поляков Алексей Евгеньевич, НПБ им. К.Д. Ушинского, снс

Добрушина Екатерина Роландовна, к.ф.н., ПСТГУ, зав.каф.

Иванова-Алленова Татьяна Юрьевна, к.ф.н., ПСТГУ, доц.

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

За 2012 год были выполнены следующие виды работ, необходимых для развития Корпуса церковнославянских текстов: (1) Разработка справочной информации к корпусу. (2) Создание терминологического словаря. (3) Выверка характеристики словоформ. (4) Разметка не разобранных программно словоформ. (5) Выверка ошибок в текстах.

Разработана справочная информация к корпусу и вывешена в рамках общей инструкции к Национальному корпусу на образовательном портале НКРЯ «Студиорум». В дальнейшем разработанная инструкция должна также быть вывешена отдельно при собственно Церковнославянском корпусе. Основные положения инструкции следующие:

Планируется, что в течение нескольких лет в разделе Исторические корпусы начнут работу следующие корпусы, связанные с историей русского языка: (1) церковнославянский; (2) XVIII века; (3) среднерусский; (4) древнерусский; (5) берестяные грамоты. В данный момент в этом разделе открыт лишь один корпус – церковнославянский, включающий тексты, созданные в XVII-XX веках. Помещение его в раздел исторических может показаться сомнительным, если исходить из того, что церковнославянский – это язык современный, живой, активно используемый определенной группой современных носителей языка, отличающийся от современного русского в первую очередь не историчностью, но сакральностью и сферой применения. Полностью разделяя эту позицию, создатели все же помещают его в раздел исторических потому, что, благодаря тщательно хранимой традиции, церковнославянский язык, несомненно, гораздо ближе к языку XVII-XVIII века, чем к современному.

Объем церковнославянского корпуса 4 700 406 слов. Он открыт в мае 2012 года в пробной версии и в течение двух лет должен быть заметно усовершенствован и снабжен собственной инструкцией.

Особенностью корпуса является наличие трех вариантов орфографии запроса: точного, упрощенного и модернизированного. Это удобно потому, что пользователь сможет найти нужное ему слово, даже если не уверен в его орфографии. Различия в типах орфографии можно увидеть, просто вызвав виртуальные клавиатуры упрощенного и модернизированного типов запросов и изучив надписи на клавишах, совмещающих буквы. Впрочем, для успешного пользования корпусом нет необходимости в этом разбираться. Варианты орфографии нужны исключительно для упрощения составления запроса. Результаты поиска всегда будут выданы в одном орфографическом варианте – классическом церковнославянском.

Чтобы сделать запрос в Церковнославянском корпусе, надо выбрать один из трех вариантов орфографии запроса (по умолчанию ставится промежуточный вариант - "упрощенный") - в зависимости от того, насколько хорошо известно правописание нужного слова. Так, если есть неуверенность, например, в выборе между "Е" и "ЯТЬ", то лучше выбрать самый простой вариант - модернизированный. В модернизированном варианте можно, например, искать слово "Бог", даже не указав "ер" на конце. Но в двух более приближенных к реальным вариантах орфографии это слово, записанное без "ера", искаться не будет. Так, слово «вѣтръ» в модернизированной орфографии будет искаться по запросам «ветр», «ветръ», «вѣтр» и «вѣтръ», а в упрощенной и точной только по «вѣтръ». Если есть желание разобраться в том, какие буквы не различаются в «упрощенном» и «модернизированном» запросах, то следует изучить набор

кнопок в виртуальной клавиатуре соответствующего режима. При точной орфографии различается 46 букв (плюс титло, которое условно трактуется как буква), в упрощенной – 39 (плюс титло), а в модернизированной – 33. В модернизированной, например, совмещены на одной кнопке буквы «еєѣ», которым в упрощенной соответствуют две кнопки – «еє» и «ѣ», а в точной – три. Набор слова в поисковой графе проще осуществлять в виртуальной клавиатуре.

Разработано содержание словарных статей терминологического словаря, поясняющего специализированные термины, использованные в разметке текстов. Термины в основном представляют собой терминологию из области славянской церковной гимнографии, например, акафист, икос, кондак, кукулий, тропарь, харетизм и др. Словарь будет вывешен в открытый доступ в 2013 году. Образец словарной статьи:

***«АКА'ФИСТ** М. (Греч. ακατηιστος, «неседальный», ακατηιστος ηψμνος, «неседальное песнопение, гимн»). Гимнографический текст со специфическими элементами структуры. /// Акафист представляет собой восхваление, как правило, адресованное Спасителю, Богоматери, Святым; оформляется по определенным жанровым правилам; начинается с кукулия, далее следуют попеременно чередующиеся икосы и кондаки, в сумме составляющие 24 строфы; каждый икос содержит 12 харетизмов; последний кондак обычно представляет собой молитвенное обращение к прославляемому; во многих изданиях первая строфа часто ошибочно названа кондаком, а не кукулием; русские акафисты, как правило, написаны по модели хронологически первого из акафистов Акафиста Пресвятой Богородице. Этот гимн, составленный, по мнению многих исследователей, в VII в. служит образцом структуры позднейших акафистов. Из всех акафистов только акафист Пресвятой Богородице по уставу входит в богослужение»*

В ходе работы над церковнославянским корпусом в 2013 году были проанализированы словоформ из рабочего словаря корпуса, в ходе этой работы были выверены грамматические характеристики словоформ, приписанные компьютерными методами, и устранены многие ошибки, возникшие из-за неправильной трактовки форм; проанализирована примерно третья часть списка словоформ, состоящего из 90000 единиц. Также были приписаны леммы и грамматические характеристики словоформам, оставшимся неразобранными в результате программного анализа; проанализировано 320 словоформ. Наконец, была проанализирована часть текстов, составляющих базу корпуса, и в них были выправлены грубые ошибки набора, в основном представляющие собой вкрапление технических комментариев типа примечаний в церковнославянский текст; обработана примерно четвертая часть текстов.

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 1

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.) *нет*

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

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

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

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

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

Ведется работа над развитием и коррекцией корпуса церковнославянского языка, с 2011 года функционирующего в пилотной версии в составе Национального корпуса русского языка. За 2012 год разработано справочное обеспечение Корпуса, а именно инструкция для пользователя и справочник по сложной терминологии, использованной в разметке. Корпус представляет собою набор текстов, объемом в 4,6 миллиона словоупотреблений, снабженный характеризующей тексты разметкой и поиском по грамматическим характеристикам. Тексты требуют выверки ошибок набора, и за 2012 год обработана и выверена примерно четверть текстов. Работа корпуса опирается на электронный словник из 150 000 словоформ, изъятых из обрабатываемого корпуса текстов, сведенных к начальным леммам и охарактеризованных грамматически. Этот словник создан автоматически на основе анализа ручной обработки некоторой части словоформ и требует проверки грамматических трактовок анализатора и правки ошибок; в течение 2012 года с этой точки зрения обработана примерно треть разобранных словоформ. Также продолжена работа над теми словоформами, которые не удалось разобрать автоматическими методами, так как они требуют научной работы специалиста; обработана треть списка подобных форм и для большей части из них найдена удовлетворительная трактовка.

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

За 2012 год были выполнены следующие виды работ: (1) Разработана справочная информация к корпусу и вывешена в рамках общей инструкции к Национальному корпусу на образовательном портале НКРЯ «Студиорум». (2) Разработано содержание словарных статей терминологического словаря, поясняющего специализированные термины, использованные в разметке текстов. (3) Выверены грамматические характеристики словоформ, приписанные компьютерными методами. (4) Приписаны леммы и грамматические характеристики словоформам, оставшимся неразобранными в результате программного анализа. (5) Проанализирована с точки зрения поиска ошибок набора четверть текстов, составляющих базу корпуса.

14. Краткий финансовый отчет за 2012 г. (основные статьи расходов по проекту, сумма) *350000*

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

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

В 2011 году был создан пилотный рабочий вариант корпуса церковнославянского языка, работающий и уже активно используемый исследователями, но требующий значительной доработки и развития. Для такого развития корпуса за 2013 год планируется сделать следующее:

– Отредактировать и вывесить в свободный доступ при Национальном корпусе русского языка в виде автономной инструкции к Церковнославянскому корпусу созданные в 2012 году справочные материалы, описывающие особенности корпуса и состав текстов и облегчающие пользователям работу с корпусом.

– Отредактировать и вывесить в свободный доступ при Национальном корпусе русского языка созданный в 2012 году краткий словарем, поясняющий специализированные термины, использованные в разметке текстов.

– Выверить приписанные компьютерными методами грамматические характеристики второй трети словоформ и устранить ошибки, возникшие из-за неправильной трактовки форм.

– Приписать леммы и грамматические характеристики словоформам, оставшимся неразобранными в результате программного анализа (около 3000 неразобранных словоформ).

– Произвести путем анализа необработанных слов и их значений в текстах поиск ошибок, возникших при наборе текстов, и выправить эти ошибки (не менее 300 текстов).

– Проанализировать значение лемм, определенных разметчиками как имена нарицательные, и приписать тем из них, перевод которых на русский язык не ясен без привлечения специальных знаний, краткое толкование, доступное пользователю вместе с грамматической характеристикой при нажатии на выбранное слово на странице результатов поиска (от 1000 лемм).

Подпись руководителя проекта А.А. Плетнева

Форма 2
Планируемое содержание работ на 2013 г.

№	Название проекта	Организация-исполнитель и учрежде-ния-соисполнтели	Руководитель проекта (+ кол-во ис-полнителей)	Запрашиваемый объем финансирования на 2013 г. (тыс. руб.)	Ожидаемые в 2013 г. результаты
2.1.	Развитие корпуса церковнослаянских текстов	ИРЯ РАН	К.ф.н. Плетнева А.А.		1. Разметка толкованиями семантически проблемных нарицательных лемм (от 1 000 лемм). 2. Выверка характеристики словоформ (от 30 000 проверенных словоформ). 3. Разметка не разобранных программно словоформ (от 3 000 словоформ). 4. Выверка ошибок в текстах (не менее 300 текстов). 5. Редактирование и размещение в свободном доступе Национального Корпуса инструкции и словаря терминов, разработанных в 2012 году.