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

<table>
<tr><td colspan="2">Номер и название раздела программы
3. Создание и развитие корпусных ресурсов по языкам народов России</td></tr>
<tr><td colspan="2">Название проекта
Создание корпусов на языках народов Северной Сибири</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Гусев Валентин Юрьевич, к.ф.н., с.н.с. Института языкознания РАН</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города) почтовый адрес руководителя проекта</td><td>valentin.gusev@iling-ran.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования

ФГБУН
Институт языкознания
(ИЯз РАН)</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Алпатов Владимир Михайлович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации — адресата финансирования
Клезович Ирина Ивановна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
тел. (495)690-35-85, факс (495)690-05-28
iling@iling-ran.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="15">Исполнители (ФИО, уч. ст.)</td><td>Андреева Т.Е., к.ф.н., ИГИиПМНС СО РАН,</td></tr>
<tr><td>Стручков К.Н., к.ф.н., ИГИиПМНС СО РАН,</td></tr>
<tr><td>Захарова Н.Е., ИГИиПМНС СО РАН;</td></tr>
<tr><td>Нестерова Е.В., к.ф.н., ИГИиПМНС СО РАН;</td></tr>
<tr><td>Шарина С.И., к.ф.н., ИГИиПМНС СО РАН,</td></tr>
<tr><td>Прокопьева П. Е., к.пед.наук, ИГИиПМНС СО РАН</td></tr>
<tr><td>Прокопьева А. Е., к.ф.н., ИГИиПМНС СО РАН</td></tr>
<tr><td>Лукина М. П., ИГИиПМНС СО РАН</td></tr>
<tr><td>Попова Н. И., к.ф.н., ИГИиПМНС СО РАН</td></tr>
<tr><td>Данилова Н. И., д.ф.н., ИГИиПМНС СО РАН</td></tr>
<tr><td>Д. Ф. Н., к.ф.н., ИГИиПМНС СО РАН</td></tr>
<tr><td>Урманчиева А. Ю., к.н.ф., ИЯз РАН</td></tr>
<tr><td>Шлуинский А. Б., к.ф.н., ИЯз РАН</td></tr>
<tr><td>Коломацкий Д. И., к.ф.н., ИЯз РАН</td></tr>
<tr><td>Волков О. А., МГУ</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. Вяч. Вс. Иванов

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

«___»_____2012 г.

1. Название направления

Создание и развитие корпусных ресурсов по языкам народов России

2. Название проекта

Создание корпусов на языках народов Северной Сибири

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

Гусев Валентин Юрьевич, к.ф.н., с.н.с. Института языкознания РАН

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

Андреева Тамара Егоровна, к.ф.н., зам. директора по науке ИГИиПМНС СО РАН,
Стручков К.Н., к.ф.н., н.с. ИГИиПМНС СО РАН,
Захарова Н.Е., н.с. ИГИиПМНС СО РАН;
Нестерова Е.В., к.ф.н., н.с. ИГИиПМНС СО РАН;
Шарина Сардана Ивановна, к.ф.н., зав. сектором ИГИиПМНС СО РАН,
Прокопьева Прасковья Егоровна, к.пед.наук, с.н.с., зав. сектором ИГИиПМНС СО РАН
Прокопьева Александра Егоровна, к.ф.н., м.н.с. ИГИиПМНС СО РАН
Лукина Маргарита Петровна, н.с. ИГИиПМНС СО РАН
Попова Наталья Иннокентьевна, к.ф.н., зам.директора ИГИиПМНС СО РАН
Данилова Надежда Ивановна, д.ф.н., зав. сектором ИГИиПМНС СО РАН
Дьячковский Федор Николаевич, к.ф.н., зав. сектором ИГИиПМНС СО РАН
Урманчиева Анна Юрьевна, к.н.ф., н.с. Института языкознания РАН
Шлуинский Андрей Болеславович, к.ф.н., н.с. Института языкознания РАН
Коломацкий Дмитрий Игоревич, к.ф.н., м.н.с. Института языкознания РАН
Волков Олег Александрович, студент Московского государственного университета

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

В 2012 году в рамках проекта шла работа над корпусами следующих языков Северной Сибири: ительменского, якутского, эвенского, эвенкийского, юкагирского и энецкого. Работа проводилась совместно силами сотрудников Института языкознания РАН (головная организация), Института гуманитарных исследований и проблем малочисленных народов Севера (Якутск) и МГУ.

Работа для разных языков велась по двум направлениям. Для относительно крупных языков, на которых есть значительное число текстового материала, готовились материалы для автоматической обработки — то есть тексты, выровненные по предложениям, и машинные словари, которые будут использованы для автоматического анализа текста. Для малых языков выбрана другая стратегия — глоссирование в полуавтоматическом режиме с использованием программ FLEx и Toolbox.

В результате выполнения проекта получены следующие результаты:

Ительменский язык — тексты из книги А. П. Володина «Ительменский язык» (Л., 1976), глоссированные в программе FLEx, общим объемом 1000 предложений (около 7000 слов).

Эвенский язык — выровненные параллельные эпические тексты «Иркэнмэл» и «Нелтэк» общим объемом 9000 словоупотреблений и машинный словарь на основе эвенско-русского словаря В. А. и М. Е. Роббеков (Новосибирск, 2005, ок. 14000 словарных статей).

Эвенкийский язык — машинный словарь на основе эвенкийско-русского словаря А. Н. Мыреевой (Новосибирск, 2004, ок. 30000 словарных статей).

Юкагирский язык — глоссированные эпические тексты «Легенда об Эдилвее» (тундровый диалект, около 2400 слов) и «Петр Бэрбэкин» (лесной диалект, около 2000 слов).

Якутский язык — эпическое сказание «Кыыс Дебелийэ» (36000 слов) и машинный словарь на основе орфографического словаря якутского языка («Сахалыы таба суруйуу тылдьыта»; Якутск, 2002, около 45000 словарных статей).

Энецкий язык — полевые записи 2009-2011 годов и архивные материалы, глоссированные тексты в программе Toolbox, около 24000 словоупотреблений.

Для всех языков, кроме юкагирского, это единственные на данный момент корпуса, выложенные в открытом доступе.

Все результаты выкладывались на корпусной интернет-ресурс, который находится на сайте Института языкознания РАН: corpora.iling-ran.ru. В 2012 году была запущена его полнофункциональная версия.

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)
Составлены или увеличены корпуса текстов на языках Северной Сибири: ительменском (ранее публиковавшиеся тексты, 7000 словоформ), якутском (сказание «Кыыс Дебелийэ, 36000 словоформ), эвенском (сказание «Иркэнмэл» и «Нелтэк», 9000 словоформ), юкагирском (сказания 4400 словоформы), энецком (24000 словоформ). Созданы машинные словари для автоматических анализаторов на эвенском (14000 слов), эвенкийском (30000 слов), якутском (45000 слов) языках. Создаваемые корпуса выкладываются в свободный доступ на сайте www.corpora.iling-ran.ru.

13. Наиболее значимый научный результат проекта (5-6 строк для сводного отчета в Президиум РАН)
Составлены или увеличены корпуса текстов на языках Северной Сибири: ительменском (7000 словоформ), якутском (36000 словоформ), эвенском (9000 словоформ), юкагирском тундровом и лесном (сказания «Легенда об Эдилвее» и «Петр Бэрбэкин», 4400 словоформ), энецком (полевые записи 2009-2011 годов и архивные материалы, 24000 словоформ). Созданы машинные словари для автоматических анализаторов на эвенском (14000 слов), эвенкийском (30000 слов), якутском (45000 слов) языках. Создаваемые корпуса выкладываются в свободный доступ на сайте www.corpora.iling-ran.ru.

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 планируется продолжать работу по тем же языкам. Необходимо будет написать морфологические анализаторы для якутского, эвенского и эвенкийского языков. Помимо этого, на 2013 год ставятся следующие задачи:

    Ительменский язык — увеличение корпуса на 7000 словоформ

    Юкагирский язык — увеличение корпуса на 10000 словоформ на обоих диалектах

    Якутский язык — подготовка текстов для автоматического анализа (20000 слов), выверка и снятие омонимии (20000 слов)

    Эвенский язык — подготовка текстов для автоматического анализа (5000 слов), выверка и снятие омонимии (5000 слов)

    Эвенкийский язык — подготовка текстов для автоматического анализа (10000 слов), выверка и снятие омонимии (10000 слов)

    Энецкий язык — увеличение корпуса на 25000 словоформ на лесном диалекте.

    Подпись руководителя проекта В. Гусев

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
	Создание корпусов на языках народов Северной Сибири	ИЯз РАН ИГИиПМНС СО РАН	Гусев В. Ю. (+ 15 чел.)		— морфологические анализаторы для эвенского, якутского и эвенкийского языков — ительменский язык — увеличение корпуса на 7000 словоформ, ручное глоссирование — юкагирский язык — увеличение корпуса на 10000 словоформ на обоих диалектах, ручное глоссирование — якутский язык — подготовка текстов для автоматического анализа (20000 слов), выверка и снятие омонимии (20000 слов) — эвенский язык — подготовка текстов для автоматического анализа (5000 слов), выверка и снятие омонимии (5000 слов) — эвенкийский язык — подготовка текстов для автоматического анализа (10000 слов), выверка и снятие омонимии (10000 слов) — энецкий язык — увеличение корпуса на 25000 словоформ на лесном диалекте, ручное глоссирование.