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 3. Создание и развитие корпусных ресурсов по языкам народов России</td></tr>
<tr><td colspan="2">Название проекта
Развитие и пополнение электронного корпуса текстов на языках малочисленных народов Сибири (на материалах ненецкого, телеутского, шорского и эвенкийского языков)</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Шаховцов Кирилл Геннадиевич, к.и.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>kyrill@iea.ras.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования
Федеральное государственное бюджетное учреждение науки Ордена Дружбы народов Институт этнологии и антропологии им. Н.Н. Миклухо-Маклая Российской академии наук (ИЭА РАН)</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Тишков Валерий Александрович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Плетнева Наталья Сергеевна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
119991 Москва, Ленинский пр-т, 32а
тел.: (495) 938-17-47, факс: (495) 938-06-00
info@iea.ras.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="3">Исполнители (ФИО, уч. ст.)</td><td>Мамонтова Надежда Александровна</td></tr>
<tr><td>Терехина Александра Николаевна</td></tr>
<tr><td>Функ Дмитрий Анатольевич, д.и.н., проф.</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 языкам народов России
2. Название проекта
   Развитие и пополнение электронного корпуса текстов на языках малочисленных народов
3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)
   Шаховцов Кирилл Геннадиевич, к.и.н., научный сотрудник ИЭА РАН
4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)
   Мамонтова Надежда Александровна, ИЭА РАН, аспирантка
   Терехина Александра Николаевна, ИЭА РАН, аспирантка
   Функ Дмитрий Анатольевич, д.и.н., проф., ИЭА РАН, зав. отделом Севера и Сибири
5. Содержание фактически проделанной работы, полученные результаты, их новизна, научная и практическая значимость, актуальность, соответствие мировому научному уровню (до 5 стр.)

Структура таблицы, лежащей в основе проекта базы данных, была доработана с учетом предполагаемого семикратного увеличения объема корпуса и значительного расширения функциональности. Программное обеспечение веб-интерфейса корпуса адаптировано к использованию системы интернационализации GNU gettext. Выполнен черновой перевод на английский и шорский языки.

Начата работа по подготовке словарей основ и аффиксов для эвенкийского и шорского языков (в формате hunspell) для реализации полнотекстового (с учетом морфологии) поиска в нормализованных текстах на этих языках средствами СУБД PostgreSQL.

Поскольку помимо фольклорных в корпус включены тексты, охраняемые законодательством об авторском праве (газетные и проза шорских и эвенкийских писателей), было реализовано разграничение по объему примеров

в поисковой выдаче: из охраняемых текстов выдаются отдельные предложения без возможности просмотра полных текстов (аналогично подходу, используемому в НКРЯ), для фольклорных произведений в настоящее сохранен неограниченный доступ к полным текстам, аудиозаписям и сканированным изображениям страниц.

С текущим состоянием корпуса можно ознакомиться по адресу http://corpora.iea.ras.ru/new_corpora/. В последствие новый интерфейс полностью заменит разработанный в рамках программы 2011 г.

Эвенкийские тексты

Объем эвенкийского подкорпуса увеличен на 25 тыс. словоупотреблений. В подкорпус устной речи включены тексты, записанные в 2010, 2011 и 2012 гг. в Эвенкийском муниципальном районе (образцы эвенкийского фольклора и рассказы). За год удалось расшифровать все собранные ранее тексты, а также осуществить их перевод на русский и частично на английский языки. Кроме того, в Корпус были включены некоторые фольклорные тексты, собранные Г.М. Василевич, из книги «Исторический фольклор эвенков», три предания о Чинанае из книги Г.И. Варламовой «Сказания восточных эвенков», а также два текста, записанных Г.В. Шахирзяновой (сотрудницей краеведческого музея п. Туры Эвенкийского муниципального района) от П.А. Удыгира (п. Эконда) в 1980-х гг и выпущенных в виде CD. Подкорпус письменной речи пополнился за счет художественных произведений А.Н. Немтушкина и Г.Н. Калитина на литературном языке и Декларации прав человека.

Большая часть текстов из подкорпуса устной речи была отредактирована при участии носителей языка и снабжена надстрочными знаками долготы. Последнее крайне важно, т.к. зачастую фольклорные тексты на эвенкийском языке публикуются без долгот, несмотря на наличие у них смыслоразличительной функции.

Шорские тексты

Объем шорского подкорпуса увеличен за счет текстов ранее неизвестных эпических текстов на 35844 словоупотреблений в оригинальной части и 25587 словоупотреблений в нормализованной части. Впервые опубликованы (размещены в Корпусе) такие эпические тексты из репертуара шорского сказителя В.Е. Таннагашева (1932–2007) как «Қусқун қараттыг Алып-Қусқун», «Ачазы қулатпа туңмазы қулат», «Ақ-Пилек» и «Қырық қулаш сынныг қара сараттыг Алып-Қарачын». Около 20 тыс. словоупотреблений (19553) переведены на русский язык. Это тексты шести эпических сказаний: «Күңнӱ көрчең Күн-Көөк», «Қырық қулаш сынныг қара сараттыг Қан-Мерген», «Ай қараттыг Қара-Қан», «Сыбазын-Оолақ», «Талашқа ч□рген Алтын-Торғу» и «Қара-Қан».

Кроме этого, была начата работа по оцифровке, набору, нормализации произведений шорских писателей 1930—1990-х гг. (в частности, произведения Ф.С. Чиспиякова) и газетных статей 1930-х гг. Из заявленного на весь

срок выполнения проекта подкорпуса письменной речи объемом ок. 40 тыс. словоупотреблений объем выполненной в этом году работы составил более 30% (ок. 15 тыс. словоупотреблений): это художественная проза и тексты религиозного содержания.

Велась работа по подготовке к включению в Корпус двух текстов героического эпоса из записей Д.А. Функа 1983-85 гг. на казасском варианте нижнемрасского говора мрасского диалекта шорского языка. Начерно были расшифрованы эпические сказания «Алтын-Оңус» и «Чеқатый-алып» из репертуара М.Е. Токмагашевой (1908-1990). Объем оригинальных расшифрованных текстов составляет ок. 12 тыс. словоупотреблений.

В общей сложности объем шорского подкорпуса увеличен за год примерно на 110 тыс. словоупотреблений.

Телеутские тексты

Объем текстов на телеутском языке увеличен на 20 тыс. словоупотреблений за счет ввода в корпус полевых записей Д.А. Функа 1980—1990-х гг., а также опубликованных фольклорных произведений и газетных текстов. Около трети всех текстов переведены на русский язык.

Ненецкие тексты

Для корпуса ненецкого языка были отсканированы, технически отредактированы или набраны тексты из сборника «Ненецие ”лаханако”» («Ненецкие сказки»), сост. Н.М. Янгасова, Салехард, 2007 г. В сборник вошли фольклорные тексты, записанные в Ямало-Ненецком автономном округе: материалы газеты «Няръяна Ӈэрм», сборника «Северные россыпи» (1962 г.) и материалы частных библиотек, а так же ранее не опубликованные произведения устного народного творчества. Аналогичная работа проведена со сборником мифов, сказок, исторических, преданий «Ненецкий фольклор» из серии «Фольклор народов Таймыра», выпуск 5, Красноярск, 1995 г.

Обработаны для помещения в Корпус статьи из ненецкой газеты «Няръяна Ӈэрм», выпускаемой в Ямало-Ненецком автономном округе. Во время экспедиционного выезда в Таймырский Долгано-Ненецкий муниципальный район были собраны 15 фольклорных текстов на таймырском говоре ненецкого языка от сказительницы Яр Софьи Эйновны. Данные тексты и материалы из Ямало-Ненецкого автономно округа находятся в процессе расшифровки и обработки с привлечением носителей языка. Объем ненецкого корпуса в настоящее время составляет 75 тыс. словоупотреблений.

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

6.1. количество монографий 1

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.)

Шорский героический эпос. Том 3: Сыбазын-Олак. Выспоренная Алтын-Торгу. Кара-Хан / Сост., подгот. к изд., статьи, пер. на рус. яз., приложения, примеч. и коммент. Д.А. Функа; сказитель В.Е. Таннагашев. Составление компакт-диска Д.А. Функа. Подготовка мастер-диска К.Г. Шаховцова. Кемерово: ООО «Примула», 2012. – 279 с.

В третьем томе серии «Шорский героический эпос» представлены три эпических сказания, ни одно из которых до сих пор не было известно в сибирской фольклористике. Все публикуемые тексты — из репертуара выдающегося шорского сказителя, Владимира Егоровича Таннагашева (1932—2007). Публикация шорских текстов и переводов на русский язык сопровождается вводными статьями, этнолингвистическими примечаниями и комментариями, а также приложениями, в том числе на CD. В статьях рассматриваются вопросы практически неизвестной традиции исполнения эпоса о смерти главного героя, анализируются проблемы из области археологии Саяно-алтайского региона (вопрос о семантике средневековых балбалов), а также ранее не исследовавшаяся проблема смены кодов в речи шорских сказителей. При подготовке текстов к изданию, написании аналитических очерков, включая текстологические примечания и комментарии, а также при составлении глоссария широко использовались возможности созданного нами Корпуса. Издание адресовано этнологам, фольклористам, археологам, студентам университетов профильных специализаций, а также широкому кругу читателей, интересующихся культурой народов Сибири, в первую очередь самим шорцам.

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

Д.А. Функ, Н.А. Мамонтова, К.Г. Шаховцов. Электронный корпус фольклорных текстов на языках малочисленных народов Сибири (на материалах шорского, телеутского и эвенкийского языков): принципы создания и структура // Мультимедийные и цифровые технологии в собирании, сохранении и изучении фольклора / Сост. В.Л. Кляус, Е.В. Миненок. Под ред. В.М. Гацака. М., 2012. С. 162-179.

Функ Д.А. Электронный корпус фольклорных текстов на языках, находящихся под угрозой исчезновения // Основные тенденции развития алтаистики в изменяющихся мировоззренческих условиях. Материалы Международной научно-практической конференции, посвященной 1150-летию российской государственности, 90-летию Ойротской автономной области, 60-летию Научно-исследовательского института алтаистики им. С.С. Суразакова. Горно-Алтайск, 26-30 июня 2012 г. Часть 2 / Отв. ред. Н.М. Екеева. Горно-Алтайск, 2012. С. 145–151.

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.)

Терехина А.Н.: экспедиция в Таймырский Долгано-Ненецкий муниципальный район (г.Дудинка, район с. Носок), с 29 июня по 25 июля 2012 г. Записаны 15 фольклорных текстов на таймырском говоре.

Мамонтова Н.А.: полевое исследование в поселках Туре и Учами Эвенкийского муниципального района Красноярского края (с 25 июня по 15 июля). За время работы удалось записать 15 новых текстов, в основном фольклорных, на разных говорах илимпийского диалекта эвенкийского языка. При участии носителей эвенкийского языка были расшифрованы и переведены на русский язык все тексты, собранные в районе во время предыдущих полевых выездов в 2010 и 2011 гг. В поселке Учами Н.А. Мамонтовой удалось записать всего 2 коротких рассказа. Это обусловлено социолингвистической ситуацией – в этом поселке фактически не осталось носителей языка, основным средством коммуникации служит русский язык. Некоторые жители знают язык пассивно. В этом же поселке проведен социолингвистический опрос жителей, а также записаны несколько десятков глубинных интервью, касающихся функционирования эвенкийского языка и проблем его ревитализации.

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)

Произведено существенное пополнение корпусов (эвенкийского — на 25 тыс. словоупотреблений, ненецкого — на 75 тыс., шорского — на 110 тыс., телеутского — на 20 тыс.) за счет как ранее опубликованных, относящихся к различным жанрам художественной литературы и фольклора, так и впервые записанных текстов, а также газетных статей.

Программное обеспечение корпуса адаптировано к системе интернационализации GNU gettext, позволяющей осуществить перевод интерфейса на любое необходимое число языков. Выполнены черновые переводы на английский и шорский языки. Начата работа над грамматическими словниками и словарем основ и аффиксов для эвенкийского и шорского языков.

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

Произведено существенное пополнение всех четырех корпусов, в общей сложности на 210 тыс. словоупотреблений за счет как ранее опубликованных, так и впервые записанных текстов. Программное обеспечение адаптировано к системе интернационализации GNU gettext, позволяющей осуществить перевод интерфейса на любое необходимое число языков.

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. будет продолжен ввод текстов: на ненецком языке 90 тыс. словоупотреблений; на телеутском 40 тыс., на шорском 160 тыс., на эвенкийском 37 тыс. Планируется завершение основной работы над словарем основ и аффиксов для эвенкийского и шорского языков с предоставлением возможности полнотекстового поиска через веб-интерфейс, продолжение работы над грамматическими словниками и словарями для шорского, телеутского и ненецкого языков. Будет выполнен черновой перевод интерфейса на ненецкий язык.

Подпись руководителя проекта К.Г. Шаховцов

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
	Развитие и пополнение электронного корпуса текстов на языках малочисленных народов	иэа ран	1 + 3		Ввод текстов: на ненецком языке 90 тыс. словоупотреблений; на телеутском — 30 тыс., на шорском — 100 тыс., на эвенкийском — 37 тыс. Завершение основной работы над словарем основ и аффиксов для эвенкийского и шорского языка; перевод интерфейса на ненецкий язык.