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
3. Создание и развитие корпусных ресурсов по языкам народов России</td></tr>
<tr><td colspan="2">Название проекта
Создание корпусов миноритарных тюркских языков России</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Дыбо Анна Владимировна, д.ф.н., проф., член-корр. РАН
Соруководитель: Широбокова Н. Н., д.ф.н., ИФ СО РАН</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>adybo@mail.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования
ФГБУН
Институт языкознания Академии наук
(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
125009 Москва, Б.Кисловский пер., 1/12
(495)2903585
(495)2900528iling@iling.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>Чебодаева Л.И., к.ф.н., ИСАТ ХГУ</td></tr>
<tr><td>Кыржинакова Э.В., к.ф.н., ХакНИИЯЛИ</td></tr>
<tr><td>Боргояков В.А., к.ф.н., ХГУ</td></tr>
<tr><td></td><td>Мальцева В. С., РГГУ</td></tr>
<tr><td></td><td>Шеймович А.В., Институт языкознания РАН</td></tr>
<tr><td></td><td>Крылов Ф.С., РГГУ</td></tr>
<tr><td></td><td>Гусев В. Ю., к.ф.н., Институт языкознания РАН</td></tr>
<tr><td></td><td>Бавуу-Сюрюн М.В., к.ф.н., ТувГУ</td></tr>
<tr><td></td><td>Невская И. А., д.ф.н., ИФ СО РАН</td></tr>
<tr><td></td><td>Есипова А.В., к.ф.н., НГПИ</td></tr>
<tr><td></td><td>Сюрюн А. А., к.ф.н., ИЛИ РАН</td></tr>
<tr><td></td><td>Оскольская П. А., ИЛИ РАН</td></tr>
<tr><td>Дата сдачи отчета</td><td>Лемская В. М., к.ф.н., ТГПУ</td></tr>
<tr><td></td><td>Токмашев Д. М., к.ф.н., КГПА</td></tr>
</table>

Форма 1

УТВЕРЖДАЕМ

Координатор Программы акад. Вяч. Вс. Иванов

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

«___»_____2012 г.

1. **Название раздела, в рамках которого выполняется проект**

3. Создание и развитие корпусных ресурсов по языкам народов России

2. **Название проекта**

Создание корпусов миноритарных тюркских языков России

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

Дыбо Анна Владимировна, д.ф.н., член-корр. РАН, зав. Отдела урало-алтайских языков Института языкознания РАН

со-руководитель: Широбокова Наталья Николаевна, д.ф.н., проф., зав. Отдела языков народов Сибири Института филологии СО РАН

4. **Основные исполнители (ФИО, ученая степень, место работы, должность)**

Невская Ирина Анатольевна, д.ф.н., г.н.с. Отдела языков народов Сибири Института филологии СО РАН

Есипова Алиса Васильевна, к.ф.н., доц. Новокузнецкого государственного педагогического института

Сюрюн Аржаана Александровна, к.ф.н., н.с. Института лингвистических исследований РАН

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

Оскольская Полина Алексеевна, аспирантка Института лингвистических исследований РАН

Лемская Валерия Михайловна, ст. преп. Томского государственного педагогического университета

Шеймович Александра Валерьевна, м.н.с. Института языкознания РАН

Токмашев Д. М., к.ф.н., в.н.с. Кузбасской государственной педагогической академии

Чебодаева Лариса Ильинична, к.ф.н., зав. каф. Института Саяно-Алтайской Тюркологии Хакасского Государственного Университета
Кыржинакова Эльвира Валерьевна, к.ф.н., м.н.с. ХакНИИЯЛИ
Бавуу-Сюрюн Мира Викторовна, к.ф.н., зав. учебно-научного центра Тувинского Государственного Университета
Боргояков Владислав Александрович, к.ф.н., доц. Хакасского Государственного Университета

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

1. Продолжались работы над созданием корпусов: хакасского,

шорского, алтайского (с диалектными подкорпусами; в частности, алтай-кижи, телеутский, теленгитский, тубаларский, чалканский).

- В параллельный хакасско-русский корпус введены:

а) полный текст хакасского героического эпоса «Ай-Хуучин» (по изданию: Хакасский героический эпос: Ай-Хуучин / Запись и подгот. текста, пер., вступ. ст., примеч. и коммент., прил. В.Е. Майногашевой. – Новосибирск: Наука, 1997); объемом 5 а.л., 28509 хакасских словоформ, 8000 фраз.

б) два сборника рассказов Л.Чебодаевой объемом 12000 и 14500 словоформ соответственно.

Производится редактирование автоматической морфологической разметки.

- Производилась обработка полевых материалов по хакасским диалектам (сагайский и бельтирский) для включения их в диалектный подкорпус хакасского корпуса. По сагайскому диалекту (Казановка) обработаны тексты, собранные экспедициями РГГУ (2001, 2002 гг.) , РГГУ и ИЯз РАН (2007 г.), ок. 12 часов звучания (расшифровка, перевод, частично отглоссированы);

по бельтирскому диалекту частично обработаны тексты, собранные экспедицией ИЯз РАН (2011 г.), 3 часа звучания (расшифровка, перевод). Предполагается завершение глоссировки и вывешивание текстов в Интернет.

- Начата работа над составлением корпуса алтайского языка. В процессе экспедиции этого года (см. ниже) собраны материалы по ряду алтайских диалектов. К настоящему моменту:

а) устный текст на карлыкском говоре диалекта Алтай-кижи длиной 1,5 часа расшифрован и переведен, готов к вывешиванию в составе параллельного корпуса.

б) на теленгитском говоре (южный вариант, Кош-Агачский район) расшифрован, переведен и отглоссирован 1 текст, 55 фраз, 900 словоформ

в) на чалканском 2 текста, 72 и 23 фразы, расшифрованы, переведены и отглоссированы, также готовы к вывешиванию.

2. Усовершенствовался морфологический анализатор для тюркских языков. В настоящий момент он разбирает в среднем 90% словоформ эпического текста (это означает, что часть неразобранных словоформ следует отнести за счет диалектных грамматических и лексических явлений, пока не отраженных в грамматическом лексическом блоках парсера, созданного для литературного языка, и выявляемых в основном в результате работы с корпусом). На базе Хакасско-русского электронного словаря, разработанного в прошлом году, создан новый компьютерный словарь - Шорско-русский, объемом 22100 словарных статей. Ранее шорского словаря такого объема не существовало (имеющийся Шорско-русский, Русско-шорский словарь, Кемерово, 1993 - включает менее 5000 слов). Словарь тестируется на материалах шорского корпуса.

Хакасско-русский словарь, помимо грамматической разметки, теперь снабжен словообразовательной разметкой в особом поле; планируется в будущем ввести эту словообразовательную разметку в разметки корпуса.

Корпуса текстов на тюркских языков России:

— хакасский (150000 словоупотреблений, 15 часов звучания);

— чулымско-тюркский (36000 словоупотреблений, 10 часов звучания);

— телеутский (36000 словоупотреблений, 10 часов звучания);

— тувинский (300000 словоупотреблений, 15 часов звучания);

— шорский (90000 словоупотреблений, 15 часов звучания).

6. Общее число опубликованных в 2012 г. по проекту работ
6.1. количество монографий
6.2. количество сборников статей
6.3. количество статей 3

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.)

А.В.Шеймович, «О принципах построения автоматического морфологического анализатора для корпуса хакасского языка» // Материалы Международной конференции «К 150-летию Н.Ф.Катанова» (Респ. Хакасия, Абакан, 17–18 мая). Абакан, 201. С. 85-96;

А.В.Шеймович, «Некоторые особенности автоматического анализа морфологии хакасского языка (на материале корпуса)» // Материалы Международной конференции «Тюркомонгольские народы Центральной Азии: язык, этническая история и фольклор» (к 100-летию со дня рождения В.М.Наделяева) (Кызыл, 20–23 мая 2012). Кызыл, 2012. С. 34-48.

А.В.Дыбо, "Морфонологический анализ и внутренняя реконструкция: к истории башкирского языка" // Материалы Всероссийской конференции "Урал – Алтай: через века в будущее" (Респ. Башкортостан, 21-23 июня). Уфа, 2012. С. 23-31.

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.)

Экспедиция в Респ. Горный Алтай, рук. Дыбо А.В., Шеймович А.В., Коровина Е.В. (8-20 августа), совместно с Ин-том языка, литературы и истории им. Суразакова, Горноалтайским университетом и Отделом языков народов Сибири Института филологии СО РАН: сбор материалов по диалектам алтайского языка.

Экспедиция имела двоякую направленность: а) сбор языкового материала, б) демонстрация коллегам из Ин-та языка, литературы и истории им. Суразакова и из Горноалтайского университета, привлекаемым к работе по проекту, основных приемов полевого исследования, направленного на пополнение корпуса, в современных условиях и с современным приборным обеспечением.

Собирались прежде всего а) устные тексты для корпуса алтайского языка;

б) 200-словные списки Сводеша, применяемые для пилотной классификации говоров;
в) фонетическая программа (1500 единиц), направленная на выявление полной картины фонологических и квазифонологических противопоставлений в конкретном говоре. Эта программа была заново составлена для экспедиции применительно к диалектам Горного Алтая А.В.Дыбо и О.А.Мудраком.
Собран материал по а) карлыкскому говору диалекта Алтай-кижи (в сел. Кзыл-Озек, в совокупности 12 часов записи, устный текст длиной 1,5 часа расшифрован и переведен, готов к вывешиванию в составе параллельного корпуса;
б) по лебединскому диалекту, сел. Курмач-Байгол (в совокупности 18 часов записи);
в) по диалекту туба-кижи, сел. Усть-Пыжа, Новотроицк, Йогач (в совокупности 36 часов записей).
Особенно важными представляются результаты сбора материалов по диалекту туба-кижи. Систематически собирались материалы этого диалекта Н.А.Баскаковым в 1934-1952 гг. С тех пор диалект подвергся сильному вымыванию; но даже и материалы Баскакова не позволяют принять однозначного решения о классификационной принадлежности диалекта к северной или южной группам алтайских диалектов. Экспедиция МГУ 2006 года (в результате которой появился сборник статей "Тубаларские этюды", М., 2008), будучи нацеленной прежде всего на получение материалов для лингвистической типологии, не занималась четким отбором информантов-носителей именно тубаларского диалекта, поэтому представленные в сборнике данные часто скорее относятся либо к литературному алтайскому, либо к чресполосному с "правобережными" тубаларами кумандинскому. Нам в этом сезоне удалось найти и опросить десятерых носителей тубаларского диалекта с левого берега Бии; в результате значительно прояснились наши представления об этом идиоме. Все собранные материалы после полной расшифровки войдут в корпус алтайского языка, диалектный подкорпус.

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)

Объем параллельного литературного хакасско-русского корпуса увеличен на 55 тыс. словоформ; в качестве текстового материала введены эпические тексты и образцы новейшей литературы на хакасском языке с переводами. Начато создание диалектного подкорпуса хакасского корпуса (звуковые файлы с пофразовой расшифровкой и переводом, ок. 15 часов звучания). Начата систематическая работа над диалектным подкорпусом алтайского корпуса. В результате экспедиции по проекту собраны важные материалы по исчезающему тубаларскому диалекту алтайского языка. Электронный хакасско-русский словарь объемом 22500 слов снабжен словообразовательной разметкой. Создан и тестируется электронный шорско-русский словарь объемом 22500 слов. По итогам работы опубликованы статьи.

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

Объем параллельного литературного хакасско-русского корпуса увеличен на 55 тыс. словоформ; начата работа над диалектным подкорпусом. В результате экспедиции по проекту собраны материалы по исчезающему тубаларскому диалекту алтайского языка. Электронный хакасско-русский словарь объемом 22500 слов снабжен словообразовательной разметкой. Создан электронный шорско-русский словарь объемом 22500 слов.

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 продолжать работу над

а) хакасским корпусом: увеличивать объем литературного параллельного хакасско-русского корпуса за счет ввода произведений классической хакасской литературы и газетных текстов (ориентировочно до 150000 словоупотреблений); продолжать расшифровку собранных диалектных текстов; собрать новые тексты на диалектах хакасского языка.

б) алтайским корпусом: диалектным подкорпусом, обработать имеющиеся звуковые файлы (теленгитские ок. 10 часов звучания, чалканские ок. 10 ч. звучания, тубаларские ок. 12 часов звучания), сделать расшифровку и глоссировку.

в) шорским корпусом: с помощью автоматического анализатора сделать морфологическую разметку эпических шорских текстов в объеме 30000 словоупотреблений.

г) увеличить объем корпуса чулымского языка на 12000 словоупотреблений;

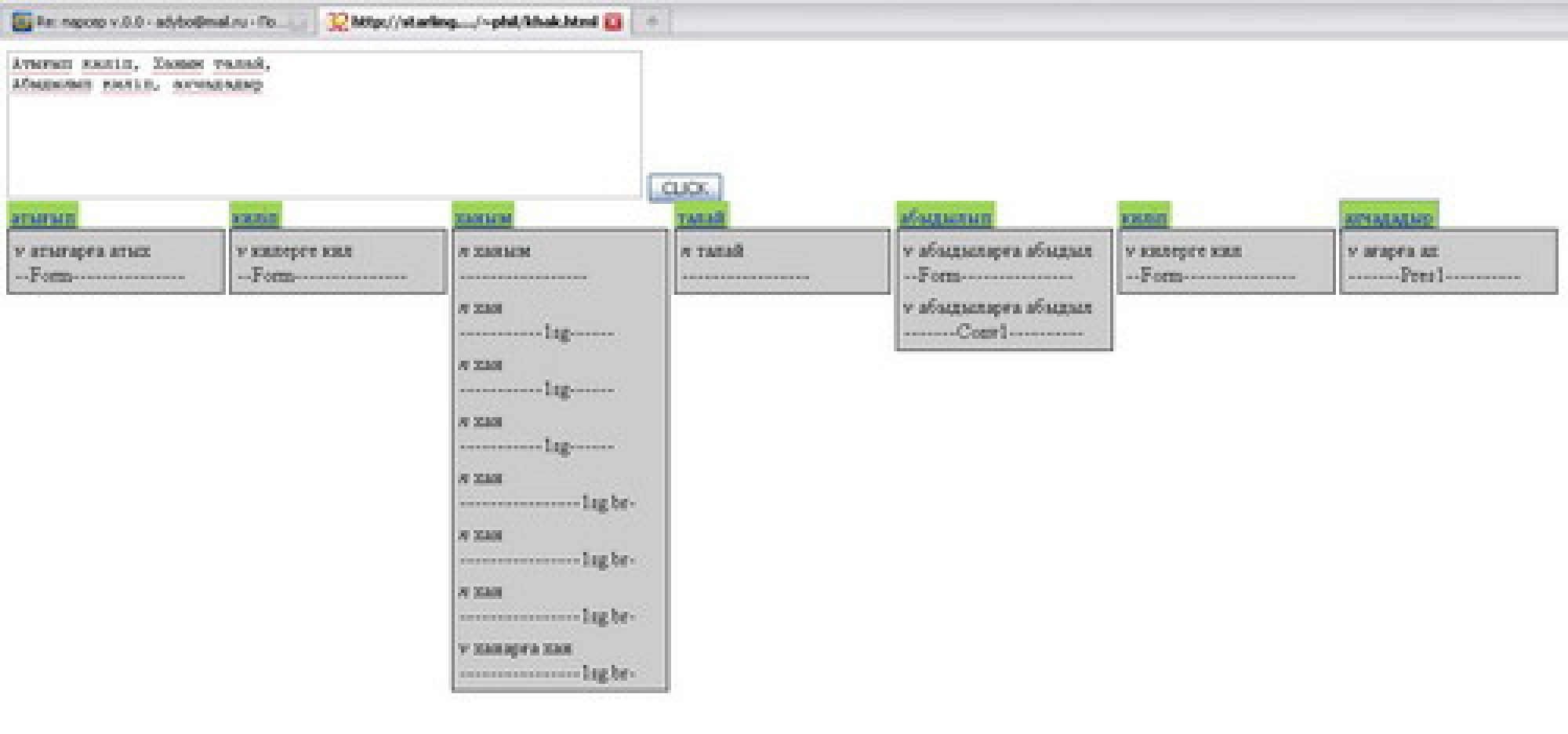
д) увеличить тувинский корпус - обработка, включающая грамматическую разметку и перевод, а также последующее выкладывание на сайт, тувинского героического эпоса (по изданию: Тувинские героические сказания: Хунан-Кара. Боктуг-Кириш, Бора-Шэлей / Сост., вступ. ст., подгот. текста, подстр. пер., коммент. и словари С.М. Орус-оол; Подгот. тувинских текстов под сред.

Д.А. Монгуша; пер., коммент. к пер. А.В. Кудиярова. — Новосибирск: Наука, 1997).

Подпись руководителя проекта А.В. Дыбо

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 миноритар-ных тюркских языков России	Институт языкознания РАН, Институт фи-лологии СО РАН	Дыбо А.В. 16		**2013 г. —** Корпуса текстов на тюрк-ских языков России: — хакасский (50000 словоупотребле-ний); — чулымско-тюркский (12000 слово-употреблений); — алтайские диалекты (12000 слово-употреблений); — тувинский (100000 словоупотреб-лений); — шорский (30000 словоупотребле-ний).


Re: парсер v.0.0 - ablybo@mail.ru - Po...
http://st.urfeng...../~phil/khak.html
Атыгын килэ, Хымэм тыпай,
Абыдыкыя килэ, ачгыкалэр
CLICK
атыгын
у атыгара атых
--Form----------
кэкэт
у кэкэрге кэкэт
--Form----------
хымэм
я хымэм

я хэкэ
----------1tg------
я хэкэ
----------1tg------
я хэкэ
----------1tg------
я хэкэ
----------1tg be-
я хэкэ
----------1tg be-
я хэкэ
----------1tg be-
у хэкэгэ хэкэ
----------1tg be-
тыпай
я тыпай

абыдыкыгын
у абыдыкыгэ абыдыкы
--Form----------
у абыдыкыгэ абыдыкы
-------Com1---------
кэкэт
у кэкэрге кэкэт
--Form----------
ачгыкалэр
у ачгэгэ ач
-------Pen1---------