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
Корпуса литературных языков Дагестана: аварский и даргинский языки</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Дмитрий Сергеевич Ганенков, к.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>d.ganenkov@gmail.com

125009 Москва Б. Кисловский пер. 1</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования

Федеральное государственное бюджетное учреждение науки «Институт языкознания РАН»

ИЯз РАН</td><td>ФИО (полностью) руководителя организации – адресата финансирования

Владимир Михайлович Алпатов</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

(495) 6917875, (495) 6900528
6917875@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="3">Исполнители (ФИО, уч. ст.)</td><td>Н.К. Богомолова, к.ф.н., Институт языкознания РАН</td></tr>
<tr><td>М.А. Даниэль, к.ф.н., МГУ им. М.В. Ломоносова</td></tr>
<tr><td>Подпись руководителя проекта:</td></tr>
<tr><td>Дата сдачи отчета 20.11.2012</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
3. Создание и развитие корпусных ресурсов по языкам народов России.

2. Название проекта

Корпуса литературных языков Дагестана: аварский и даргинский языки

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

Ганенков Дмитрий Сергеевич, к.ф.н., с.н.с. отдела кавказских языков Института языкознания РАН

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

Н.К. Богомолова, к.ф.н., Институт языкознания РАН, н.с. отдела кавказских языков Института языкознания РАН

М.А. Даниэль, к.ф.н., кафедра ТиПЛ филологического факультета МГУ им. М.В. Ломоносова

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 проделана следующая работа и получены следующие результаты:

- автоматическое распознавание сканированных текстов — 1000 усл.печ.л.;
- ручная корректура автоматически распознанных текстов — 50 усл.печ.л. (около 400000 словоупотреблений);
- сбор газетных текстов на даргинском литературном языке, опубликованных на интернет-портале www.zamana.etnosmi.ru — 250000 словоупотреблений;
- создание грамматического словаря и морфологических парадигм даргинского языка;

- морфологическое аннотирование электронных текстов (лемма, словоизменительные и словообразовательные характеристики словоформы, перевод леммы на русский язык) — 510000 словоупотреблений;
- подготовка метатекстовой разметки аннотированных текстов — 25 текстов;
- индексация корпуса аннотированных текстов в поисковом движке — 510000 словоупотреблений (25 текстов);
- размещение тестового корпуса на интернет-ресурсе http://dag-languages.org/corpora/dargwa.

Основным результатом работы по проекту является тестовый корпус даргинского литературного языка, объемом 510000 словоупотреблений и обеспечена возможность его регулярного пополнения при наличии текстов в электронном виде.

Корпус даргинского языка создается в рамках проекта по созданию морфологически аннотированных корпусов литературных языков Дагестана.

Глубина морфологической разметки (а именно: полная морфологическая аннотация каждой словоформы) даргинского корпуса превосходит имеющиеся корпуса многих крупных языков Европы.

Появление такого корпуса выводит работу по исследованию даргинского языка на совершенно новый уровень, позволяя за мгновения сформировать представительную выборку примеров употребления изучаемых языковых единиц (лексем, грамматических форм, синтаксических конструкций, фразеологизмов) путем поиска по лексеме, русскому переводу лексемы, грамматическим характеристикам отдельных лексем, а так же разнообразным сочетаниям как признаков одной лексемы, так и сочетаниям разных лексем и их признаков. Корпус значительно облегчит и сделает более удобным изучение грамматической семантики и синтаксиса даргинского языка, ранее возможное только в полевой работе с информантами или библиотечной работе с художественной литературой.

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 0

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)
В 2012 г. разработан грамматический словарь и морфологические парадигмы, необходимые для автоматической обработки текстов на даргинском литературном языке. Подготовлен к обработке парсером значительный объем художественной литературы на даргинском языке (около 400000 словоупотреблений). Собран корпус газетных текстов на даргинском языке (объемом около 250000 словоупотреблений). При помощи программы-парсера подготовлен к размещению на интернет-сервере (произведена морфологическая и метатекстовая разметка, индексация в поисковом движке, размещение в интернете) тестового корпуса даргинского литературного языка объемом 510000 словоупотреблений.

13. Наиболее значимый научный результат проекта (5-6 строк для сводного отчета в Президиум РАН)
В ходе проекта разработан грамматический словарь и морфологические парадигмы, необходимые для автоматической обработки текстов на даргинском литературном языке. При помощи программы-парсера подготовлен (произведена морфологическая и метатекстовая разметка, индексация в поисковом движке) и размещен в интернете тестовый корпус даргинского литературного языка объемом около 510000 словоупотреблений .

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 создать тестовый аннотированный корпус аварского литературного языка. В течение советского (начиная с 20-х гг. XX-го века) и постсоветского периода на аварском языке был опубликован значительный по объему корпус прозаических текстов различных жанров (рассказы, повести, романы, сказки, анекдоты, публицистика, духовная литература). Развитие современных компьютерных технологий позволяет использовать тексты для детального исследования грамматики и лексики этих языков. В настоящее

время подавляющее большинство текстов на аварском литературном языке представлено в виде книг, небольшая часть (тексты из газеты на аварском языке «ХIакъикъат») представлена в электронном виде на портале www.etnosmi.ru. Общий объем опубликованных текстов на аварском языке, по нашим оценкам, составляет не менее 15 млн. словоупотреблений. Конечной целью является создание корпусов аварского литературного языка объемом 7-10 млн. словоупотреблений.

В 2013 г. планируется осуществить первый этап создания такого корпуса. В рамках этого проекта ставятся следующие задачи

1) подготовка текстов для корпуса — в связи с тем, что большинство текстов на аварском языке представлено в виде книг, требуется значительная работа по их переводу в электронный формат (сканирование, автоматическое распознавание, ручная корректура автоматически распознанных текстов);

2) подготовка грамматического словаря и морфологических таблиц аварского языка;

3) автоматическая разметка электронных текстов при помощи морфологического анализатора.

Аннотация корпуса будет производиться при помощи программы UniParser в формате, аналогичном формату созданного ранее корпуса лезгинского литературного языка. Каждая словоформа в корпусе будет снабжена информацией о лексеме, части речи, словоизменительных грамматических категориях, а также переводом лексемы. Корпус аварского литературного языка будет базироваться на поисковом движке EANC и позволит осуществлять различные виды поиска по лексеме, переводу лексемы, грамматическим и семантическим характеристикам как отдельных лексем, так и их сочетаний.

Ожидаемый результат работы в 2013 г. — создание тестового корпуса аварского литературного языка объемом около 500000 словоупотреблений и его размещение в свободном доступе на веб-сайте.

Подпись руководителя проекта Д.С. Ганенков

Форма 2

Планируемое содержание работ на 2013 г.

№	Название проекта	Организация-исполнитель и учреждения-соисполнители	Руководитель проекта (+ кол-во исполнителей)	Запрашиваемый объем финансирования на 2013 г. (тыс. руб.)	Ожидаемые в 2013 г. результаты
1	Корпуса литературных языков Дагестана: аварский и даргинский языки	Институт языкознания РАН	к.ф.н Д.С. Ганенков (+ 2 исполнителя)		2013 г. — тестовый корпус аварского литературного языка (объем 500000 словоупотреблений)