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
Электронные корпуса новописьменных лезгинских языков:
агульский и удинский</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Майсак Тимур Анатольевич, к.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>timur.maisak@gmail.com</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования
Институт языкознания РАН
(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
(495) 6903585
(495) 6900528
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="3">Исполнители (ФИО, уч. ст.)</td><td>Ганенков Дмитрий Сергеевич, к.ф.н., ИЯз РАН</td></tr>
<tr><td>Ландер Юрий Александрович, к.ф.н., ГУ ВШЭ / ИВ РАН</td></tr>
<tr><td>Мерданова Солмаз Рамазановна, д.ф.н., МГППУ / ИЯз РАН</td></tr>
<tr><td>Дата сдачи отчета
19.11.2012</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
   3. Создание и развитие корпусных ресурсов по языкам народов России

2. Название проекта
   Корпуса новописьменных лезгинских языков: агульский и удинский

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)
   Майсак Тимур Анатольевич, к.ф.н., с.н.с. отдела кавказских языков ИЯз РАН

4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)
   Ганенков Дмитрий Сергеевич, к.ф.н., с.н.с. отдела кавказских языков ИЯз РАН

   Ландер Юрий Александрович, к.ф.н., доцент факультета филологии ГУ ВШЭ, н.с. отдела языков ИВ РАН

   Мерданова Солмаз Рамазановна, д.ф.н., профессор МГППУ, в.н.с. отдела кавказских языков ИЯз РАН

5. Содержание фактически проделанной работы, полученные результаты, их новизна, научная и практическая значимость, актуальность, соответствие мировому научному уровню (до 5 стр.)
   Проект состоит из двух частей, объединяемых общей целью — это создание электронных корпусов для двух новописьменных языков лезгинской группы нахско-дагестанской семьи, агульского и удинского.

Этап 2012 года был посвящен развитию того «задела», который уже имелся по агульскому корпусу, а для удинского языка этап стал начальным и был в первую очередь посвящен подготовке текстовых материалов.

1) АГУЛЬСКИЙ. Основной массив текстов на литературном агульском языке составляют статьи в районной газете «Вести Агула» (http://agul.etnosmi.ru/) и перевод Евангелия от Луки, выпущенный Институтом перевода Библии в 2005 году. В 2011 г. в рамках корпусной программы была создана пилотная версия агульского корпуса на платформе *SIL Fieldworks Language Explorer*. В 2012 г. произошло пополнение текстовых ресурсов за счет обработки новых статей из газеты «Вести Агула» (около 16 тыс. слов). Поскольку на сайте газеты тексты приводятся в требующем коррекции виде, в ходе обработки они были нормализованы путем устранения опечаток и технических ошибок набора. Еще одним, ранее не запланированным, источником текстов стало издание «Свод памятников фольклора народов Дагестана: в 20-ти томах», первые два тома которого (Том 1. Сказки о животных; Том 2. Волшебные сказки) вышли в 2011 году и стали доступны в 2012 году. В этих томах опубликованы пять сказок на агульском языке (около 4 тыс. слов), с русским переводом. В связи с большим количеством опечаток и орфографических ошибок, эти тексты также были нормализованы для корпуса. Тем самым, общий прирост текстов составил около 20 тыс. слов. Помимо этого, ранее разобранный в *SIL Fieldworks* перевод Евангелия от Луки (20,5 тыс. слов) в этом году был снабжен русским переводом. Для этого был выбран не синодальный, а более современный русский перевод Нового Завета, выполненный В. Н. Кузнецовой (БИБЛИЯ. Современный русский перевод. М.: Рос. Библейское о-во, 2011. — 1408 с.); предисловие и словарь трудных слов переведены Т. А. Майсаком. Попутно проходило пополнение словаря и морфологических правил разбора, снятие омонимии. Версия агульского Евангелий от Луки с переводом готовится к размещению на сайте http://corpora.iling-ran.ru.

2) УДИНСКИЙ. Основная задача, которая решалась на данном этапе — создание собственно материала для последующего автоматического анализа, т.е. электронной библиотеки удинских литературных текстов: их сканирование и распознавание (либо набор) и выверка по бумажному изданию. Электронные версии были подготовлены таким образом для большей части изданий современной удинской прозы, поэзии и фольклора на ниджском диалекте, а именно:

- Кечаари Ж. (сост.) Нана очъал: Шеирхо, гьекйаьтхо, драма. Баку, 1996.
- Aydınov Y. A., Keçaari J. A. Əlifba (TIetIir). Bəkü, 1996.
- Aydınov Y. A., Keçaari J. A. Udin muz (udi dili) 3. Bakı, 1996.
- Keçaari Ƙ. Orayin. Bakı, 2001.
- Keçaari Ƙ. Buruxmux. Gəncə, 2003.
- Dabakov V. V. (сост.) Udiğoy folklor: Nağılxo. Legendoox. Астрахань, 2007.

Указанные публикации представляют три варианта удинской орфографии, бытовавших в разное время: а) кириллическая, б) созданная на ее основе латинизированная (с использованием азербайджанских букв и с опорой на диграфы) и в) новая латиница, избегающая диграфы и активно применяющая диакритики. В целях единообразия все тексты были переведены в в кириллицу. Общий объем текстов составил около 60 тыс. слов.

Помимо этого, начата работа над написанием правил морфологического анализа для удинского языка (ниджского диалекта) в рамках системы автоматического парсинга *UniParser* (разработана Т. А. Архангельским).

6. Общее число опубликованных в 2012 г. по проекту работ
   нет
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.)
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)
Для агульского корпуса: произошло пополнение текстовых ресурсов за счет добавления новых статей из газеты «Вести Агула» за 2012 год (ок. 16 тыс. слов) и сказок из первых двух томов издания «Свод памятников фольклора народов Дагестана» (ок. 4 тыс. слов) По сравнению с исходным видом на сайте газеты (http://agul.etnosmi.ru/) тексты были нормализованы путем устранения опечаток и технических ошибок набора. Ранее разобранный в *SIL Fieldworks* перевод Евангелия от Луки (20,5 тыс. слов) снабжен русским переводом; пополнен словарь и морфологические правила. Для удинского корпуса: создана основная часть электронной библиотеки (сканирование и распознавание либо набор и выверка по бумажному изданию) для дальнейшего автоматического анализа, общий объем текстов составил около 60 тыс. слов. Написана часть правил морфологического анализа для ниджского диалекта (в рамках системы *UniParser*).

13. Наиболее значимый научный результат проекта (5-6 строк для сводного отчета в Президиум РАН)
Пополнены текстовые ресурсы агульского корпуса (газетные статьи — ок. 16 тыс. слов, сказок — ок. 4 тыс. слов). К ранее размеченному агульскому переводу Евангелия от Луки добавлен русский перевод; пополнены словарь и морфологические правила. Создана электронная библиотека удинских текстов для дальнейшего автоматического анализа (ок. 60 тыс. слов); написана часть правил морфологического анализа.

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).

Основные направления работы в 2013 году для удинского корпуса:

* подготовка дополнения к библиотеке текстов для морфологического анализа (около 40 тыс. слов);

* дальнейшая разработка и завершение удинского парсера;

* анализ подготовленной библиотеки текстов (имеющиеся 60 тыс. + новые 40 тыс. слов) при помощи разработанного парсера, пополнение словаря, исправление найденных ошибок.

Основные направления работы в 2013 году для агульского корпуса:

* подготовка дополнения к библиотеке текстов для морфологического анализа (газетные статьи, 15—20 тыс. слов);

* конвертация в формат XML агульских диалектных текстов, отглоссированных ранее в MS Word (не менее 45 тыс. слов)

Подпись руководителя проекта

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
	Корпуса новописьменных лезгинских языков: агульский и удинский	Институт языкознания РАН	Т. А. Майсак, к.ф.н.		Доработка удинского парсера, подготовка дополнительного объема удинских текстов для парсинга (40 тыс. слов), подготовка и парсинг дополнения к агульскому корпусу (до 20 тыс. слов), конвертация агульских диалектных текстов (45 тыс. слов).