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 4. Создание и развитие корпусных ресурсов по языкам мира.</td></tr>
<tr><td colspan="2">Название проекта
Восточноамянский национальный корпус: развитие ресурса</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Вячеслав Всеволодович Иванов, д.ф.н., академик РАН</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>ivanov@ucla.edu
(</td></tr>
<tr><td>Полное и краткое название организации – адресата финансирования

Институт славяноведения РАН
ИнСлав РАН</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Никифоров Константин Владимирович</td></tr>
<tr><td></td><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Боченова Наталья Владимировна</td></tr>
<tr><td></td><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
119334 Москва, Ленинский пр., д. 32-А
Телефон: +7 (495) 938-17-80</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>Исполнители (ФИО, уч. ст.)</td><td>М.А. Даниэль, к.ф.н., ОТиПЛ МГУ им. Ломоносова, НИУ ВШЭ</td></tr>
<tr><td></td><td>Т.А. Архангельский, НИУ ВШЭ
А.А. Печеный, студент фил. факультета МГУ им. Ломоносова</td></tr>
<tr><td></td><td>В.Г. Хуршудян (консультант)
С.В. Рубаков (консультант)</td></tr>
<tr><td>Дата сдачи отчета</td><td>20.11.12</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 мира

2. Название проекта
Восточноармянский национальный корпус: развитие ресурса

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)
Вячеслав Всеволодович Иванов, д.ф.н., ак. РАН

4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)
М.А. Даниэль, к.ф.н., НИУ ВШЭ, профессор; ОТиПЛ МГУ им. Ломоносова, доцент

Т.А. Архангельский, НИУ ВШЭ, преподаватель

А.А. Печеный, МГУ им. Ломоносова (студент филологического факультета)

консультанты:

С.В. Рубаков, программист компании Яндекс (главный программист системы www.eanc.net)

В.Г. Хуршудян, преподаватель INALCO (Париж) (разработчик ресурса www.eanc.net)

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

В рамках проекта развивается Восточноармянский национальный корпус (www.eanc.net) и его поисковый механизм, являющейся важным компонентом разных проектов данного направления программы (албанский национальный корпус, осетинский корпус, лезгинский корпус, даргинский корпус, лакский корпус и проч.). За прошедший год существенным образом переработано отображение выдачи. В результате корпуса средних и малых языков могут отображаться в виде строк с поморфемным глоссированием, то есть с выравниванием словоформ с их морфологическими разборами в стандартном представлении, используемом в типологических источниках (ср., например, Leipzig Glossing Rules 2007-). Такое отображение, с одной стороны, соответствует международным научным стандартам; с другой, впервые поставлена задача применить глоссированную выдачу к корпусам такого объема. Для решения этой задачи потребовалось модифицировать структуру используемых в системе баз данных, модуля выдачи и индексатора.

С точки зрения развития самого проекта, произошло важное пополнение ресурса. В корпус добавлен подкорпус диалектных текстов, собранных в районах Арцваберда, Гусана и Шенавана. Общий объем корпусов – около 300,000 словоупотреблений (правда, лишь несколько более половины составляют собственно диалектные токены). Диалектные тексты представлены интервью арменистов-диалектологов с носителями соответствующих диалектов, которые были собраны и обработаны специально для Восточноармянского национального корпуса. Индексация осуществлена с учетом специфики текстов – индексируются только реплики носителей диалекта, поиск возможен как по диалектной лемме, так и по ее литературному эквиваленту (в тех случаях, когда они различны). Диалектные тексты выделены в специальный подкорпус, поиск по которому осуществляется отдельно от поиска по основному корпусу.

Исходным материалом для работы послужили заранее собранные и оцифрованные интервью в виде документов MS Word, которые с помощью автоматической процедуры были разделены на файлы с чистым текстом и таблицы с метаинформацией. При размещении этих текстов в корпусе особую сложность представляла их грамматическая разметка. Поскольку большинство словоформ в этих текстах отличаются от своих литературных эквивалентов (или вовсе их не имеют), использовать имеющийся морфологический парсер, применяемый нами для разметки литературных текстов, не представлялось возможным. При этом вносить изменения в алгоритм работы парсера и его словари, чтобы добавить в него возможность размечать диалектные тексты, было бы край-

не неэффективной стратегией. Во-первых, это потребовало бы кардинальной переработки словаря и словоизменительных таблиц; во-вторых, эти преобразования пришлось бы производить отдельно и независимо для каждого из диалектов; в-третьих, объём диалектных корпусов достаточно мал, что даёт возможность его ручной обработки.

Для разметки текстов была применена следующая стратегия. Для каждого диалекта был составлен полный список словоформ. Эти списки были размечены парсером для литературного языка, что позволило получить разборы для тех (относительно немногочисленных) словоформ, которые в диалекте имеют ту же форму, что и в литературном языке. После этого оставшиеся словоформы были размечены вручную с указанием их диалектных лемм и литературных эквивалентов в тех случаях, где это было возможно.

Участниками проекта были созданы специальный модуль, позволяющий использовать полученные таблицы с разборами словоформ для разметки текстов, получая на вход чистые диалектные тексты и превращая их в размеченные тексты формата EANC. Частью этого модуля является система проверки таблицы разборов на корректность (поскольку она размечалась руками, в ней неизбежен некоторый процент ошибок). Размеченные тексты было решено поместить в отдельную от литературного корпуса базу данных и выделить для поиска отдельный интерфейс, параллельный основному. Диалектный интерфейс в основном совпадает по функциональности с основным, но отличается в некоторых деталях (поиск и по литературной, и по диалектной леммам; иная система транслитерации – в диалектах присутствуют фонемы, отсутствующие в литературном языке). В настоящее время данный интерфейс проходит тестирование и будет доступен по ссылке с основной страницы eanc.net к 1 декабря.

Одновременно происходит работа по оптимизации поискового механизма сервера.

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

нет

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.)

нет

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

нет

9. Список книг, сданных в печать или поданных на издательские гранты (указать ожидаемое время издания, объем в а.л.)

нет

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)

В рамках проекта осуществляется развитие Восточноармянского национального корпуса (www.eanc.net) *и его поискового движка. Корпус пополнен диалектными текстами: около 150,000 словоупотреблений интервью с носителями армянских диалектов. Добавление диалектных текстов потребовало некоторой адапатации индексатора корпуса и поиска (поиск по диалектным и нормализованным леммам). Сделаны дополнительные улучшения в поисковых алгоритмах. В смысле интерфейса выдачи корпуса, важное изменение заключается в том, что теперь корпус может отображать не квазиглоссированную выдачу, а собственно тексты, снабженные морфологическими глоссами; таким образом, поисковый движок* eanc *можно более эффективно использовать для работы с разработанными в рамках Программы корпусами средних и малых языков. Соответствующие изменения внесены в разрабатываемый в связке с Программой парсер UniParser.*

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

Добавлены корпуса трех диалектов общим объемом около 150,000 диалектных словоупотреблений. Индексатор и интерфейс корпуса переработаны таким образом, чтобы можно было отображать тексты на малых и средних языках, снабженные поморфемным глоссированием.

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).

    Движок еапс является инфраструктурным компонентом Программы, на которую опирается большинство проектов данного направления. В будущем году мы планируем, опираясь на изменения и доработки поискового движка еапс, адаптировать его и перенести на него корпуса малых и средних языков, разрабатываемые в рамках программы РАН.

    Предполагается работа по улучшению грамматического словаря корпуса (добавление новых элементов, исправление ошибок и неоднозначностей).

    Технические работы включают перенос серверов корпуса к новому провайдеру (включая оплату работ по установке, настройке и хостингу).

Подпись руководителя проекта Вяч. Вс. Иванов

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
	Восточноармянский национальный корпус: развитие ресурса		Вяч.Вс.Иванов (+3)		Перенос сервера к новому провайдеру, адаптация движка EANC под корпуса новых средних и малых языков