

Титульный лист отчета о работе в 2012 г.  
по Программе фундаментальных исследований Президиума РАН  
«Корпусная лингвистика»

Номер и название направления Программы <b>Направление 4. Создание и развитие корпусных ресурсов по языкам мира.</b>	
Название проекта <b>Албанский национальный корпус</b>	
Научный руководитель проекта (ФИО полностью, уч. ст.) <b>Д.ф.н. Русаков Александр Юрьевич</b>	
Е-mail, телефон, факс (с кодом города) почтовый адрес руководителя проекта	<b>ayurusakov@gmail.com</b>
Полное и краткое название организации – адресата финансирования  Институт лингвистических исследований Российской Академии Наук ИЛИ РАН	ФИО (полностью) руководителя организации – адресата финансирования <b>Казанский Николай Николаевич</b>
	ФИО (полностью) главного бухгалтера организации – адресата финансирования <b>Белозёрова Надежда Павловна</b>
	Телефон, факс (с кодом города), Е-mail организации – адресата финансирования тел., факс: (812)3284611 <b>iliran@mail.ru</b>
Год начала – год окончания проекта	<b>2012—2014</b>
Объем финансирования, полученного в 2012 г.	Объем финансирования, запрашиваемый на 2013 г.
Исполнители (ФИО, уч. ст.)	<b>Морозова Мария Сергеевна, магистр</b>
	<b>Домосилецкая, Марина Валентиновна, к.ф.н.</b>
	<b>Коноваленко Анна Вадимовна, магистр</b>
	<b>Каминская Лариса Николаевна, к.ф.н.</b>
Дата сдачи отчета 20.11.12	Подпись руководителя проекта:

Координатор Программы

акад. Вяч. Вс. Иванов

Координатор Программы

чл-корр. РАН В.А. Плунгян

«\_\_\_»\_\_\_\_\_2012 г.

1. Название направления: Направление 4. Создание и развитие корпусных ресурсов по языкам мира.

2. Название проекта: Албанский национальный корпус.

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы): Русаков Александр Юрьевич, д.ф.н., вед.н.сотр., ИЛИ РАН

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

Морозова Мария Сергеевна, магистр, аспирант, ИЛИ РАН

Домосилецкая, Марина Валентновна, к.ф.н., с.н.с., ИЛИ РАН

Коноваленко Анна Вадимовна, магистр, аспирант ИЛИ РАН

Каминская Лариса Николаевна, к.ф.н., доцент, СПбГУ

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

Работа по проекту «Национальный албанский корпус» была начата в 2011 году в рамках программы фундаментальных исследований Президиума РАН «Корпусная лингвистика». В результате проведенной в 2011 г. работы была осуществлена дигитализация текстов для пробного корпуса (ок. 500000 словоформ), разработана система грамматической разметки текстов, создан пробный грамматический словарь наиболее частотной лексики, разработан программа морфологического анализа. В ноябре 2012 года пробный вариант корпуса стал доступен онлайн.

В 2012 году работа над корпусом была продолжена по заданным в 2011 году направлениям:

- Была продолжена дигитализация текстов, написанных на албанском литературном языке. В настоящее время общий объем введенных текстов составляет 3,5 млн. словоформ. В жанровом плане представлены тексты современной албанской художественной литературы, эссеистика, газетные тексты, перевод книг Нового и Ветхого завета. В настоящее время к работе

над дигитализацией текстов привлечены албанские коллеги (Приштина, Косово).

- Продолжается совершенствование системы грамматической разметки. В ходе этого процесса возникла необходимость решения ряда проблем как общего характера, так и определяемых некоторыми конкретными грамматическими особенностями албанского языка. К числу таких проблем относятся грамматическая омонимия (весьма характерная для албанской именной и глагольной систем), проблема трактовки именных и глагольных аналитических форм (для албанского характерно использование как служебных слов, так и – особенно широко – клитик, причем последние в некоторых грамматических контекстах выступают как постфиксы и инфиксы), проблема характеристики многочисленных вариативных форм, появление которых обусловлено поздним становлением литературной нормы албанского языка. Представляется, что углубленное изучение этих проблем способствует не только совершенствованию грамматической разметки корпуса, но и лучшему пониманию некоторых сложных моментов албанского грамматического описания.

- Пополняется грамматический словарь. В настоящее время лемматизировано около 7000 единиц (3166 существительных; 1982 глагола; 1646 наречий; 148 предлогов, союзов).

Надо отметить, что результаты работы над корпусом обсуждались на ряде международных конференций (Международный семинар по албанскому языку и культуре, Приштина, Косово, август 2012; Международная филологическая конференция, Санкт-Петербург, СПбГУ, март 2012; конференция, посвященная столетию со дня рождения А.В.Десницкой, сентябрь 2012, Санкт-Петербург, ИЛИ РАН) и вызвали большой интерес албанских коллег.

*Новизна* представленных результатов определяется тем, что до сих пор для албанского – одного из европейских языков, имеющего более 8 млн. носителей – до сих пор не имелось представительного электронного корпуса с грамматической разметкой.

*Научная значимость* работы заключается в том, что в случае осуществления проекта откроются совершенно новые возможности для изучения албанского языка, прежде всего, его грамматического строя. Надо отметить, что в плане углубленного грамматического описания с использованием современных лингвистических методов албанский язык находится в достаточно неблагоприятном положении. Результаты работы над корпусом могут быть использованы в процессе преподавания албанского языка, в этом заключается *практическая значимость* представленной работы.

Использование технологий примененных при создании уже существующих и успешно работающих корпусов (НКРЯ, ВАНК)

свидетельствует о *соответствии* проведенной работы *мировому научному уровню*.

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

6.1. количество монографий – 0, не планировалось

6.2. количество сборников статей – 0, не планировалось

6.3. количество статей – 1 (менее 1 п.л.).

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.)

Экспедиции в рамках проекта на данном этапе не планировались и не проводились.

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 результате проведенной в 2012 году работы общий объем дигитализированных текстов достиг 3,5 миллионов словоупотреблений, расширено жанровое разнообразие текстов (современная албанская художественная литература, эссеистика, газетные тексты, перевод книг Нового и Ветхого завета). Продолжалось совершенствование системы грамматической разметки. В процессе решения находятся такие важные проблемы как трактовка грамматической омонимии (характерной для албанской именной и глагольной систем), трактовка именных и глагольных аналитических форм, проблема единообразного представления многочисленных вариативных форм, появление которых обусловлено поздним становлением литературной нормы албанского языка. Пополнился грамматический словарь, насчитывающий в настоящее время 7000 единиц.

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

Увеличение объема дигитализированных текстов до 3,5 миллионов единиц, увеличение грамматического словаря до 7000 единиц, совершенствование системы грамматической разметки.

14. Краткий финансовый отчет за 2012 г. (основные статьи расходов по проекту, сумма)

14.1. Заработная плата с начислениями (ФИО, сумма):

Русаков Александр Юрьевич – 104409 рублей

Домосилецкая Марина Валентиновна – 30000 рублей

14.2. Договоры гражданско-правового характера с физическими лицами (ФИО, сумма, тема договора):

Морозова Мария Сергеевна – 53763 рубля 44 коп. (составление грамматического словаря албанского языка).

14.3. Расходы на экспедиции и командировки:

Русаков А.Ю.

Место: Приштина, Косово;

Срок: август 2012;

цель командирования: участие в научной конференции с докладом по результатам работы над корпусом;

общая сумма: 35000 рублей

суточные: 5000 рублей

проезд: 30000 рублей

14.4. Услуги связи – 5000

14.5. Услуги сторонних организаций (название организации, тема договора) – Издательство «Нестор-История», 35000, редакционно-издательская работа над сборником, включающим исследования по корпусу (планируется на 2013 год).

14.6. Приобретение оборудования (наименование, цена) – сканер, два съемных диска, 15000 р.

14.7. Приобретение расходных материалов – 15000 рублей

15. Запрашиваемый на 2013 г. объем финансирования, с кратким обоснованием расходов: 400 тысяч рублей. В 2013 году планируется увеличить объем дигитализированных текстов до 10 миллионов словоупотреблений, закончить создание основного грамматического словаря в объеме 30000 единиц, начать работу по снятию грамматической омонимии. Предполагается участие в работе ряда конференций (Санкт-Петербург, Приштина, Тирана) с докладами о ходе работы над корпусом, а также написание и публикация статей. Планируются также расходы на приобретение материалов и оборудования.

16. Краткое обоснование научных работ на 2013 г., ожидаемые результаты – до 1 стр. и заполненная форма 2 (см. ниже).

В 2013 году планируется увеличить объем дигитализированных текстов до 10 миллионов словоупотреблений, закончить создание основного грамматического словаря в объеме 30000 единиц, начать работу по снятию грамматической омонимии. Предполагается участие в работе ряда конференций (Санкт-Петербург, Приштина, Тирана) с докладами о ходе работы над корпусом, а также написание ряда статей.

Подпись руководителя проекта

А.Ю. Русаков

**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
	Национальный албанский корпус	ИЛИ РАН	Д.ф.н. А. Ю. Русаков + четыре исполнителя	400	В 2013 году планируется увеличить объем дигитализированных текстов до 10 миллионов словоупотреблений, закончить создание основного грамматического словаря в объеме 30000 единиц, начать работу по снятию грамматической омонимии. Предполагается участие в работе ряда конференций (Санкт-Петербург, Приштина, Тирана) с докладами о ходе работы над корпусом, а также написание ряда статей.