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
4. Создание и развитие корпусных ресурсов по языкам мира</td></tr>
<tr><td colspan="2">Название проекта
Корпус новогреческого языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
чл.-корр. РАН д. ф. н. Евгений Васильевич Головко</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>evggolovko@yandex.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования

Институт лингвистических исследований Российской Академии Наук,
ИЛИ РАН</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Николай Николаевич Казанский</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Надежда Павловна Белозёрова</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
тел. (812)3282551, факс: (812)3284611
iliran@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="5">Исполнители (ФИО, уч. ст.)</td><td>Елоева Фатима Абисаловна, д. ф. н., СПбГУ</td></tr>
<tr><td>Русаков Александр Юрьевич, д. ф. н., ИЛИ РАН</td></tr>
<tr><td>Кисилиер Максим Львович, к. ф. н., ИЛИ РАН</td></tr>
<tr><td>Архангельский Тимофей Александрович, НИУ ВШЭ</td></tr>
<tr><td>Панов Владимир Александрович, к. ф. н., ИЯ РАН</td></tr>
<tr><td>Дата сдачи отчета
20.11.2012</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: **4. Создание и развитие корпусных ресурсов по языкам мира**

2. Название проекта: **Корпус новогреческого языка**

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы): **чл.-корр. РАН д. ф. н. Евгений Васильевич Головко**

4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы):
Елоева Фатима Абисаловна, д. ф. н., СПбГУ
Русаков Александр Юрьевич, д. ф. н., ИЛИ РАН
Кисилиер Максим Львович, к. ф. н., ИЛИ РАН
Архангельский Тимофей Александрович, НИУ ВШЭ
Панов Владимир Александрович, к. ф. н., ИЯ РАН

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

Основным результатом работы стала пилотная версия корпуса новогреческого языка с морфологической разметкой.

Создание электронного корпуса языка, снабжённого грамматической разметкой, включает в себя несколько последовательных этапов: уточнение набора грамматических категорий языка; построение списка словоизменительных типов, содержащих соответствие между показателями и выражаемыми ими грамматическими значениями; создание грамматического словаря; выбор текстов для включения в корпус; сбор текстов в электронном виде и перевод их в нужный формат; разметка текстов с помощью автоматического морфологического парсера и, наконец, размещение полученного размеченно-

го корпуса в интернете с помощью поисковой платформы. Для создания пилотной версии необходимо было в той или иной мере решить все эти задачи.

Первой частью работы была выработка решения относительно набора граммем современного греческого языка; в первую очередь это касается глагольной системы. Основную проблему составляли два фактора. Во-первых, в традиционных греческих грамматиках видо-временные глагольные формы рассматриваются в рамках одной парадигмы с общим названием «время», категориями которой являются настоящее время, аорист, имперфект и т. п. Однако нами было установлено, что намного логичнее и удобнее с точки зрения широты поисковых возможностей корпуса представить эту категорию в виде комбинации собственно времени (например, прошедшее и настоящее) и аспекта (перфектив и имперфектив). Тем не менее, поскольку очевидно, что некоторым пользователям будет удобнее работать с традиционными пометами, было принято решение встроить в поисковую систему, кроме аспекта и времени sensu stricto, также традиционные обозначения «аорист» и «имперфект», которые при составлении запроса будут автоматически переводиться в комбинацию времени и аспекта. Второй проблемой являются аналитические формы (например, будущее время и перфект). Нами было установлено, что некоторые из них описываются довольно строгими правилами, что позволяет с высокой точностью производить их автоматическую разметку в текстах. Однако, поскольку не все аналитические конструкции позволяют такое описание, было решено придерживаться морфологического принципа разметки, т. е. отмечать только те глагольные категории, которые непосредственно выражены в глагольных словоформах. Такое решение позволит сохранить единообразие разметки и упростит пользование корпусом. Таким образом, было решено помету времени объявить бинарной (морфологически прошедшее vs. морфологически непрошедшее) и не выделять при разметке «зависимое наклонение».

Поскольку для автоматической разметки текстов необходим грамматический словарь, вторым шагом было построение списка словоизменительных категорий. Решено было начать работу с существительных и прилагательных, изменяющихся в греческом языке, согласно грамматикам, по нескольким различающимся словоизменительным моделям. В процессе построения формализованного описания этих типов были обнаружены типы, не описанные в грамматиках; особенно это относится к прилагательным, в которых значения нескольких параметров (прилагательное двух или трёх окончаний, паттерны сдвига ударения в косвенных формах, наборы окончаний, образование сравнительной степени, возможная вариативность), соединяясь в разных комбинациях, образуют более двух десятков словоизменительных типов. Кроме того, были описаны несколько непродуктивных типов словоизменения.

Составление грамматического словаря, т. е. сбор лексики и приписывание каждой лексеме словоизменительного типа, является одной из самых трудоёмких задач при создании корпуса. Для её облегчения было решено применить специальное техническое решение. Нами была разработана программа,

позволяющая извлекать списки лексем с предположениями об их словоизменительных типах из текстов на греческом языке. Принцип работы этой программы таков: она получает на вход списки всех словоизменительных типов в формализованном виде, читает все словоформы переданных ей текстов и пытается найти группы словоформ, левая часть которых совпадает, а правые части представляют собой флексии одного из словоизменительных типов. В таком случае программа считает найденную группу форм представителсями одной лексемы, а совпадающую левую часть — её основой. На основании полученного набора окончаний делается предположение о словоизменительном типе, в соответствии с которым предполагаемая основа достраивается до начальной формы и записывается в словарь вместе с указанием словоизменительного типа. Кроме этого, применялись дополнительные специфические для греческого языка способы фильтрации (например, при распознавании рода существительного программа учитывает, какие артикли перед ним стоят в текстах). Такая система позволила собрать довольно большой словарь высокого качества. В настоящий момент такая операция была проделана с существительными и прилагательными. Процент брака оказался низким, и в большинстве случаев ошибки программы заключались в том, что лексеме присваивалось несколько типов, отличающихся в одном-двух окончаниях; на качество разметки этот факт практически не оказывает влияния. После автоматического сбора словника была произведена его ручная проверка. Было проверено около 5000 лексем, добавлено некоторое количество недостающих частотных существительных и прилагательных, а также примерно 1000 слов получили переводы на английский язык. Благодаря такому подходу, нам удалось получить словник, содержащий более 20 тысяч лексем, что существенно превышает показатель, намеченный в плане на 2012 год (7 тысяч). Данный подход был признан продуктивным; его планируется продолжить для составления списка глаголов.

Следующим шагом работы было определение списка текстов и их сбор. Нами был определён примерный список текстов на димотике, которые желательно иметь в корпусе, и примерный баланс текстов по жанрам. В настоящий момент собраны тексты общим количеством почти 5 млн словоупотреблений. Тексты принадлежат к разным жанрам (художественная литература, научная литература, пресса, юриспруденция и т. п.) и написаны на димотике — современном варианте новогреческого языка (кроме одного текста на кафаревусе, включённого в корпус для исследования вопроса о том, насколько сильно должны будут различаться словари и грамматические описания димотики и кафаревусы). Тексты, собранные для пилотной версии корпуса, в большинстве случаев получены из открытых источников; основную работу по сбору текстов планируется провести на следующих этапах. Некоторые из текстов потребовали обработки: перевода в подходящий формат, исправления систематических ошибок и т. п.; для этого были созданы специальные программные средства.

Для морфологической разметки собранных текстов был использован универсальный парсер UniParser, созданный в ходе работы над проектами Кор-

пусной программы РАН в 2011 г. В парсер были добавлены некоторые возможности для разметки греческого языка, а также возможность разметки аналитических конструкций (которая не используется в греческом корпусе в настоящий момент, но, вероятно, будет использоваться в дальнейшем). Грамматический словарь был составлен в специальном формате, совместимом с парсером UniParser. Размеченные тексты записываются парсером в формате XML, близком к формату, используемому в Национальном корпусе русского языка. Использование XML позволяет использовать размеченные тексты корпуса с разными поисковыми платформами.

В качестве поисковой платформы для корпуса была использована платформа EANC, впервые использованная в Восточноармянском национальном корпусе. Эта платформа позволяет выполнять широкий спектр поисковых запросов и поддерживает корпуса больших размеров (согласно плану, общий размер греческого корпуса к концу проекта составит не менее 30 млн словоформ). Платформа EANC была адаптирована под нужды греческого корпуса (переводы названий, таблица выбора греческих грамматических значений, специальные функции вроде перевода пометы «аорист» в комбинацию «прошедшее время + перфектив» на лету и т. п.).

Созданный корпус в настоящее время проходит тестирование и будет доступен по адресу http://www.web-corpora.net/GreekCorpus/search/?interface_language=ru .

Новизна и актуальность работы определяются тем фактом, что новогреческий язык, являясь крупным европейским языком с древней письменной традицией, до сих пор не имеет общедоступного аннотированного корпуса. письменной традицией). Существующие два крупных ресурса по новогреческому языку по существу таковыми не являются: «Корпус греческих текстов» (http://greekcorpora.isll.uoa.gr/, 30 млн. словоупотреблений) вообще не содержит грамматической разметки и переводов на английский язык, а «Греческий национальный корпус» (http://corpus.ilsp.gr/) включает 47 млн. словоупотреблений, но плохо сбалансирован (2/3 корпуса составляет пресса), содержит только леммы и разметку по частям речи, а главное, не общедоступен, поскольку является платным. Наконец, оба корпуса включают в себя только тексты конца XX в., написанные на димотике.

Создаваемый в рамках данного проекта корпус новогреческого языка по уровню морфологической разметки и поисковых возможностей не уступает ведущим мировым электронным корпусам. Применённые при создании корпуса средства (в частности, автоматический сбор лексики из текстов) таже соответствуют мировому научному уровню.

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)

Создано формализованное описание словоизменения греческих существительных и прилагательных. Созданы программные средства, позволяющие автоматически собирать греческие прилагательные и существительные для грамматического словаря из текстов. С помощью данных средств был собран грамматический словарь, содержащий более 20 тысяч лексем. Наиболее частотные из них были проверены вручную; примерно 1000 лексем был приписан английский перевод. Полученные описания парадигм и лексем были записаны в формате, совместимом с морфологическим парсером UniParser, а сам парсер был доработан для разметки текстов на греческом языке. Определён предполагаемый текстовый состав и баланс текстов корпуса. Для пилотной версии корпуса собраны и обработаны тексты общим количеством около 5 млн употреблений. Собранные тексты были размечены и размещены в Интернете с помощью поисковой платформы EANC.

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

Создано описание словоизменения и грамматический словарь новогреческих существительных и прилагательных общим объёмом более 20 тысяч лексем, а также программные средства, позволяющие автоматически пополнять лексикон грамматического словаря. Собраны тексты общим количеством около 5 млн словоупотреблений. Размеченные тексты были в качестве пилотного варианта корпуса размещены в Интернете.

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 г. таковы:

1. Расширение словаря димотики. Списки существительных и прилагательных, собранные в 2012 г., должны быть полностью проверены вручную, а наиболее частотным лексемам должны быть приписаны переводы на английский язык. Кроме того, необходимо создать формализованное описание глагольного словоизменения в соответствии с выработанными в текущем году принципами и добавить в лексикон глаголы. Поскольку глагольная система в целом сложнее именной, для этого потребуется большее количество ручной работы, чем при сборе существительных, где большую часть лексикона удалось собрать автоматически.

2. Увеличение текстовой базы. Планируется увеличить объём корпуса до 15 млн словоупотреблений, соблюдая при этом намеченный график и баланс. Некоторые тексты, не существующие в электронном виде, будет необходимо сканировать и распознавать.

3. Работа над словарём кафаревусы. Поскольку греческий корпус не может быть полным без текстов на втором варианте новогреческого языка — кафаревусе, — необходимо составить грамматический словарь этого варианта языка. В 2013 г. планируется начать работу над грамматическим словарём кафаревусы и выполнить тестовую разметку нескольких текстов.

Подпись руководителя проекта Е. Головко

Форма 2
Планируемое содержание работ на 2013 г.

№	Название проекта	Организация-исполнитель и учреждения-соисполнтели	Руководитель проекта (+ кол-во исполнителей)	Запрашиваемый объем финансирования на 2013 г. (тыс. руб.)	Ожидаемые в 2013 г. результаты
	Корпус новогреческого языка	ИЛИ РАН	Е. В. Головко (+5)		увеличение объёма корпуса до 15 млн. словоупотреблений (в т. ч. тексты на кафаревусе), увеличение грамматического словаря димотики не менее чем до 25 000 лексем (включая глаголы), создание начальной версии словаря кафаревусы