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
Расширение системы корпусов современных языков майя и создание корпуса языка рапануи (о. Пасхи)</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Алпатов Владимир Михайлович, член-корр. РАН</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>v-alpatov@ivran.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Наумкин Виталий Вячеславович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Батяркина Татьяна Владимировна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
107031, Москва, ул. Рождественка, д. 12
тел. (495) 621-18-84, факс (495) 623-19-09
info@ivran.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="4">Исполнители (ФИО, уч. ст.)</td><td>Виноградов И. А., Институт языкознания РАН</td></tr>
<tr><td>Давлетшин А. И., к.и.н., РГГУ</td></tr>
<tr><td>Козьмин А. В., к.ф.н., РГГУ</td></tr>
<tr><td>Коровина Е. В., РГГУ</td></tr>
<tr><td>Дата сдачи отчета</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

Создание и развитие корпусных ресурсов по языкам мира

2. Название проекта

Расширение системы корпусов современных языков майя и создание корпуса языка рапануи (о. Пасхи)

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

Алпатов Владимир Михайлович, д.ф.н., проф., директор Института языкознания РАН

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

Виноградов Игорь Андреевич, м.н.с. Института языкознания РАН
Давлетшин Альберт Иршатович, к.и.н., н.с. Института восточных культур и античности РГГУ
Козьмин Артём Викторович, к.ф.н., н.с. Центра типологии и семиотики фольклора РГГУ
Коровина Евгения Владимировна, аспирант Центра компаративистики РГГУ

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

В рамках задачи расширения системы корпусов современных корпусов языков майя была проделана следующая работа: была улучшена разметка уже имеющихся отглоссированных текстов, были проделаны подготовительные работы для создания автоматического поискового механизма по текстам, были отглоссированы новые тексты на языке цоциль, были написаны программы автоматического морфологического анализа для языков чоль, чорти и цельталь, а также был отглоссирован ряд текстов на этих языках.

В рамках проекта «Создание корпусов трех языков семьи майя (цоциль, юкатекский майя, киче)», реализованного в 2011 году при поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика», были созданы три параллельных корпуса на языках цоциль, юкатекский майя и киче с переводами на испанский и / или английский языки. Эти корпуса легли в основу проекта, над которым ведётся работа в настоящее время.

В этом году в связи с планами по внедрению нового усовершенствованного поискового механизма потребовалось существенно изменить морфологическую разметку текстов. Была добавлена информация о частеречной принадлежности каждой словоформы (для каждого конкретного языка набор частей речи варьирует, но незначительно: в основном, любая вариативность подобного рода связана с выделением тех или иным отдельных классов среди неизменяемых частей речи). Там, где этого требует грамматический строй языков майя, была добавлена и информация о синтаксическом статусе части речи: например, все глаголы охарактеризованы по признаку переходности.

Для каждого из языков, которые уже входят или в скором времени войдут в разрабатываемый корпус был составлен список морфологических параметров, по которым будет возможно осуществление поиска: составлен список частей речи, возможных грамматических характеристик для каждой из них, конкретных грамматических показателей. Таким образом, проделана вся подготовительная работа, которая позволит создать и ввести в эксплуатацию поисковый механизм, с помощью которого пользователь сможет осуществлять поиск не только по тому или иному значению грамматической категории, но и по конкретному грамматическому показателю. Преобразование языковых словарей в формат, поддерживаемый морфологическим анализато-

ром и будущим поисковым механизмом, даст возможность проводить поиск и по конкретным лексемам на исследуемых языках, а также по их переводам на испанский и / или английский язык.

К уже существующим отглоссированным текстам на языке цоциль были добавлены новые общим объёмом около 20 000 слов. Основным источником стал сборник *Pérez López E., M. Hidalgo Pérez & A. Gómez Gómez (eds.). Cuentos y relatos indígenas; tomo V. México D.F.: Universidad Nacional Autónoma de México, 1994.*

В отчётном году были написаны программы автоматического морфологического анализа текстов на языках чоль, чорти и цельталь. Они позволяют с достаточно хорошей точностью автоматически анализировать морфологию каждой словоформы в заранее подготовленных и специальным образом отформатированных текстах. Сложная проблема, решённая в рамках данной задачи, – это диалектное фонологическое варьирование, приводящее к варьированию графики, характерное для языка чоль (в некоторых диалектах присутствуют фонемы и, соответственно, графические символы, которые отсутствуют в других диалектах). Эта проблема потребовала привода используемого словаря языка чоль к такому формату, по которому бы однозначно восстанавливалась соответствующая форма для каждого из диалектов. При глоссировании была введена специальная помета, указывающая на диалектные изменения в графике.

Было отглоссировано около 27 000 слов на языке чоль. Они все относятся к нарративному жанру и представляют все три основных диалекта: Тумбала, Тила и Сальто-де-Агуа. Основные источники чольских текстов:

José Alejos García. Wajalix b□ t'an. Narrativa tradicional ch'ol de Tumbalá, Chiapas. México D. F., Universidad Nacional Autónoma de México, 1988,

Victor R. Gutiérrez Martínez (ed.). Selección de cuentos. Chol. Tuxtla Gutiérrez, Gobierno del estado de Chiapas. 2001,

Relatos Choles / Albilbä tyi lakty'añ. México, Dirección General de Culturas Populares e Indígenas. 2002,

Whittaker A., V. Warkentin. Chol Texts on the Supernatural. Summer Institute of Linguistics, University of Oklahoma. 1965.

На языке чорти было отглоссировано около 15 000 словоупотреблений из двух источников:

Perez Martínez, Vitalino. Leyenda Maya Ch'orti'. Guatemala: Proyecto Lingüístico Francisco Marroquín, 1996,

[no author]. Tradición oral bilingüe de la cultura ch'orti'. Guatemala: Academia de Lenguas Mayas de Guatemala (ALMG), 2005.

На языке цельталь было в общей сложности отглоссировано более 22 000 словоупотреблений. Основными источниками цельтальских текстов послужили следующие публикации:

Relatos Tzeltales / Lo'il a'yej ta Tzeltal kóp. México, Dirección General de Culturas Populares e Indígenas. 2002,

Alarcón Estrada V., V. Esponda Jimeno, A. Gómez Gómez & E. Pérez López (eds.). Cuentos y relatos indígenas; tomo VI. México D.F.: Universidad Nacional Autónoma de México, 1997.

В ходе работ по созданию корпуса языка рапануи были получены следующие основные результаты:

Создан параллельный корпус текстов для языка рапануи. Общий объем параллельного корпуса, включающего в себя запись текстов в орфографии источника и перевод согласно источнику, составил 47 798 слов (3534 предложений).

Корпус глоссированных текстов на языке рапануи включает 41 834 слова и помимо сведений входящих в корпус параллельных текстов содержит запись текстов в нормализованной орфографии с пометами, указывающими на синтаксическое членение и морфологическое глоссирование, разработанное в ходе проекта. Морфологическое глоссирование ориентировано на лейпигские правила глоссирования, что дает возможность пользоваться корпусом не только специалистам по полинезийским языкам и языку рапануи, но и использовать его данные в типологических исследованиях.

Источниками текстов являются:

Felbermayer, Fritz. 1971. Sagen und Überlieferungen der Osterinsel. Nürnberg.

Mai ki hāpī tātou i te tai□o □e i te papa□i i to tātou □arero Rapa Nui, te hā puka 4. 1990. Valparaíso: Universidad Católica de Valparaíso and Instituto Lingüístico de Verano.

Weber, Robert L. 1988. The verbal morphology of Rapa Nui : the Polynesian language of Easter Island, and its function in narrative discourse. M.A. thesis, University of Texas at Arlington.

Помимо корпуса был создан поисковый механизм, позволяющий искать по морфологическим параметрам разметки. Корпус глоссированных

текстов, снабженный поисковым механизмом, размещен в Интернете по адресу http://newstar.rinet.ru/~kozmin/rapanui/rapanui.php

Для автоматического глоссирования текстов на языке рапануи была создана специальная программа. Эта программа включает в себя элементы синтаксического анализа, таким образом, речь идет о построении так же синтаксического анализатора рапануи, которая при некоторой модификации может быть использована для автоматического глоссирования текстов на других полинезийских языках.

Для автоматического морфологического глоссирования создан электронный словарь языка рапануи, общим объемом около 2000 полнозначных слов, в основу которого были положены следующие источники:

Comisión para Estructuración de la Lengua Rapanui. 2000. Diccionario etimológica Rapanui-Español. Valparaiso, Chile: Puntángeles Universidad de Playa Ancha Editorial.

Englert, Sebastian. 1978. Idioma rapanui: Gramática y diccionario del antiguo idioma de la Isla de Pascua. Santiago, Chile: Ediciones de la Universidad de Chile.

Fuentes, Jordi. 1960. Diccionario y gramática de la lengua de la Isla de Pascua. Santiago, Chile: Editorial Andr'es Bello.

Hernández Sallés, Arturo et al. 2001. Diccionario ilustrado: Rapa Nui, Español, Inglés, Francés. Santiago, Chile: Pehuén Editores.

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

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.)

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 рамках задачи расширения системы корпусов современных корпусов языков майя была проделана следующая работа: улучшена разметка уже имеющихся отглоссированных текстов, проделаны подготовительные работы для создания автоматического поискового механизма по текстам, отглоссированы новые тексты на языке цоциль (общим объёмом около 20 000 слов), написаны программы автоматического морфологического анализа для языков чоль, чорти и цельталь, а также отглоссирован ряд текстов на этих языках (общим объёмом около 65 000 слов).

Создан параллельный глоссированный корпус языка рапануи объемом 41 834 слова, а также параллельный корпус текстов объемов 47 798. Разработана система морфологического глоссирования для языка рапануи. Создана программа автоматического морфологического глоссирования с элементами синтаксического анализа для языка рапануи. Глоссированный корпус размещен в Интернете и снабжен поисковым механизмом для морфологических помет. Создан электронный словарь языка рапануи.

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

Общий объём лингвистических корпусов на различных языках семьи майя был увеличен примерно на 85 000 словоупотреблений, и теперь составляет около 200 000 словоупотреблений. Создан корпус языка рапануи объемом около 40 000 слов. Созданы программы морфологического анализа текстов на шести языках семьи (цоциль, киче, юкатекский майя, чоль, чорти и цельталь). Создана программа морфологического анализа для языка рапануи.

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 проведение работ по размещению лингвистических корпусов в интернете. Для этого будут задействованы услуги программистов, специализирующихся в области корпусной лингвистики. Будет создан специальный сайт, на котором будет размещена основная информация о проекте и об исследуемых языках.

На этот сайт будут выложены все готовые на данный момент корпуса. Будет произведена интеграция многофункционального поискового механизма по размеченным текстам. Таким образом, результаты уже проведённой на данный момент работы станут доступны для широкого круга пользователей, в распоряжении которых окажутся современные инструменты для проведения самостоятельных исследований в интересующих их областях лингвистики (морфология, синтаксис, грамматическая и лексическая семантика, диалектология, лексикография и т.д.).

В 2013 г. планируется значительное расширение объема корпуса рапануи. Основное внимание будет уделено также расширению электронного словаря, а также синтаксической разметки корпуса. Планируется также расширение функциональности поискового модуля корпуса, который в настоящее время уже размещен в Интернете.

Подпись руководителя проекта В.М. Алпатов

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
	Расширение системы корпусов современных языков майя и создание корпуса языка рапануи (о. Пасхи)	ИВ РАН	В. М. Алпатов (+ 4)		• создание интернет-сайта с основной информацией о проекте и об исследуемых языках семьи майя; • размещение на сайте уже готовых отглоссированных и размеченных текстов на языках майя; • написание и внедрение поискового механизма для корпуса майя; • расширение объема корпуса текстов и электронного словаря рапануи; • усовершенствование синтаксической разметки корпуса рапануи и расширение функциональности поискового модуля