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 1. Создание и развитие корпусных ресурсов по современному русскому языку</td></tr>
<tr><td colspan="2">Название проекта Корпуса и коллекции интерферированных вариантов русского языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Летучий Александр Борисович, к.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>alexander.letuchiy@gmail.com</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования
Институт русского языка им. В.В.Виноградова РАН, ИРЯ РАН</td><td>ФИО (полностью) руководителя организации – адресата финансирования Молдован Александр Михайлович</td></tr>
<tr><td>Почтовый адрес, телефон, факс (с кодом города), E-mail организации – адресата финансирования Москва, 121019, ул. Волхонка, 18/2, (495) 695-26-60 Факс: (495) 695-26-03 E-mail: irlras@mail.ru</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования Глебова Татьяна Николаевна</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>Даниэль М.А., к.ф.н., НИУ ВШЭ</td></tr>
<tr><td>Добрушина Н.Р., к.ф.н., НИУ ВШЭ</td></tr>
<tr><td>Марушкина А.С., к.ф.н., НИУ ВШЭ</td></tr>
<tr><td>Резникова Т.И., к.ф.н., НИУ ВШЭ</td></tr>
<tr><td>Дата сдачи отчета 29.11.2012</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 **Направление 1. Создание и развитие корпусных ресурсов по современному русскому языку**

2. Название проекта **Корпуса и коллекции интерферированных вариантов русского языка**

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы) **Летучий Александр Борисович, к.ф.н., н.с., Институт русского языка РАН**

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

   Даниэль Михаил Александрович, к.ф.н., Факультет филологии НИУ ВШЭ, профессор
   Добрушина Нина Роландовна, к.ф.н., Факультет филологии НИУ ВШЭ, доцент
   Марушкина Анастасия Сергеевна, к.ф.н., Факультет филологии НИУ ВШЭ, доцент
   Резникова Татьяна Исидоровна, к.ф.н., Факультет филологии НИУ ВШЭ, доцент

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

В ходе осуществления проекта в 2012 году основные задачи, поставленные на этот год, были выполнены: как и планировалось, уже размеченные тексты размещены онлайн и снабжены поисковым движком; разработано универсальное рабочее место для создания и разметки корпусов ошибок. Установлены связи с Международным Центром исследований наследуемого языка (Center for World Languages and National Heritage Language

Resource Center), созданный корпус используется специалистами при обучении русскому языку, Центр собирает тексты для пополнения корпуса.

Тестовая версия корпуса херитажных текстов запущена и доступна в Интернете по адресу: http://web corpora.net/HeritageRussian/search/?interface_language=ru

Для корпуса были взяты тексты, собранные и переданные нам Международным Центром исследований наследуемого языка (Center for World Languages and National Heritage Language Resource Center).

Тексты представляют собой эссе, ответы на вопросы и т. п., записанные от информантов, посещающих курс русского языка. Общий объём корпуса равен 22 тыс. слов, или 91 текст; помимо уже введенных в корпус текстов имеется дополнительно массив ок. 10-15 тыс. слов, который подвергся ручной разметке ошибок и служит эталонным массивом для тестирования программ. Эти тексты тоже будут помещены в корпус при ближайшей вывеске (пополнении).

Тексты были снабжены метаразметкой, содержащей, среди прочего, информацию о носителе языка, жанре текста и времени записи. Метаинформация была записана в специальном виде в начале каждого файла с текстом.

В ходе работы тексты были подвергнуты процедуре автоматической морфологической разметки и преобразованы в формат, пригодный для их вывески в Интернете. Метаинформация была отделена от текстов и записана в отдельные файлы, а сами тексты были сохранены в виде чистых текстовых файлов в кодировке UTF-8. Каждый текст затем был размечен с помощью морфологического парсера mystem, используемого в Национальном корпусе русского языка.

Из-за большого количества орфографических ошибок процент верных разборов в херитажных текстах ниже, чем в стандартных русских текстах, однако благодаря функции угадывания морфологического разбора многие неправильно записанные слова всё же получили правильные морфологические характеристики. Размеченные файлы с помощью специально написанной программы были преобразованы в формат, используемый поисковой платформой EANC. Поисковая платформа была соответствующим образом настроена (в частности, был задан список возможных грамматических помет и создан графический интерфейс, позволяющий выбрать пометы при поиске). После этого файлы были проиндексированы, а настроенная платформа и полученная база данных были выложены в Интернет.

Платформа EANC позволяет производить поиск по текстам корпуса. Поисковый запрос может содержать в себе определённую словоформу, лемму, набор грамматических помет или комбинацию этих элементов (включая дизъюнкцию, конъюнкцию и отрицание). Возможен поиск нескольких словоформ с указанием диапазона расстояний между ними и поиск слов внутри одного предложения. Поиск может быть ограничен подкорпусом — частью текстов, отобранных по значениям метапризнаков (эта возможность в херитажном корпусе будет реализована при ближайшей вывеске).

Для увеличения количества правильных разборов был составлен список слов, не распознанных морфологическим парсером, отсортированный по частотности. В тех случаях, где контекст и возможная омонимия не влияли на однозначность понимания неправильно написанного слова, наиболее частым из этих слов были приписаны исправленные варианты. Этот список будет использован при следующей вывеске корпуса, что позволит существенно увеличить количество правильных разборов.

Рабочее место разметчика, которое позволяет приписывать вручную пометы, классифицирующие ошибки в херитажных текстах, функционировало в пилотном режиме: в течение года отрабатывалась и уточнялась система помет, необходимых для разметки. В частности, введена помета аспектуальных ошибок и расклассифицированы типы калек, среди которых теперь различаются полные и частичные.

По результатам проведенной работы был сделан доклад на конференции "Маргиналии-2012: границы культуры и текста" (Касимов 2012), а также на III международном коллоквиуме по лексической типологии (Гранада 2012).

Выделен полный список лексико-синтаксических ошибок в размеченном массиве, проведен предварительный анализ их причин и сформулированы лингвистические задачи описания конкретных случаев расхождения текста с русской языковой нормой. В частности, системной ошибкой херитажных текстов является сбой в построении сочинительных конструкций, вызванный смещением (под влиянием английского языка) значения сочинительных союзов. На базе ВШЭ организован семинар по изучению грамматики херитажных текстов. Предполагается, что в 2013 году будет осуществлен ряд исследований такого рода ошибок.

Параллельно ведется работа над созданием программы автоматического поиска сочетаемостных ошибок в нестандартных (в том числе, херитажных) текстах. С этой целью херитажный корпус, не достигший пока нужного для статистических исследований объема, наращивается текстами другого рода,

также содержащими значительный процент сочетаемостных, синтаксических и лексических ошибок в русском языке.

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

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.)** Не предусмотрено

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.) Не предусмотрено

10. Экспедиции, проведенные в рамках проекта Не предусмотрено

11. Конференции, организованные в рамках проекта (название, место и сроки проведения, обсуждаемые проблемы, результаты)

III международный коллоквиум по лексической типологии (17-19 сентября, Гранада 2012)

Посвящен типологическим исследованиям в области лексики; в рамках этого коллоквиума обсуждалась проблематика калькирования, имеющая прямое отношение к типологическим задачам.

Речь шла об ошибках нестандартных говорящих (в частности, носителей русского языка, которые живет в иноязычной среде), касающихся лексического выбора, который делается под влиянием доминирующего языка. Типология возможностей выбора для таких случаев, как показало обсуждение, дает новый ракурс для решения данной проблемы.

Семинар по исследованию особенностей херитажного русского (НИУ ВШЭ)

12. Важнейшие научные результаты работы по проекту (ок. 0,5 стр. для публикации на сайте Программы)

Тестовая версия корпуса херитажных текстов (текстов носителей русского языка, которые живут в иноязычной среде и унаследовали язык от родителей) запущена и доступна в Интернете по адресу http://web-corpora.net/HeritageRussian/search/?interface_language=ru ;

Отлажено и используется разметчиками рабочее место для обработки и исследования херитажных текстов; уточнен и расширен список помет, которые используются в разметке;

В рамках III международного коллоквиума по лексической типологии (17-19 сентября, Гранада 2012) проведено обсуждение подходов к описанию грамматики ошибок нестандартных говорящих на русском языке.

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

Тестовая версия корпуса «херитажных» текстов (текстов носителей русского языка, которые живут в иноязычной среде и унаследовали язык от родителей) запущена и доступна в Интернете по адресу http://web-corpora.net/HeritageRussian/search/?interface_language=ru .

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 ходе выполнения проекта в 2013 году мы планируем осуществить следующие виды работ.

1) Подготовка публикации, отражающей результаты работы по проекту
2) Разработка программы автоматического поиска ошибок сочетаемостной природы
3) Создание представительного корпуса ошибок с метаразметкой, отражающей разные типы и источники текстов (херитажные, региональные варианты, РКИ, академическое письмо, некачественные переводы и др.)
4) Организация банка данных по текстам для Грамматики ошибок

Подпись руководителя проекта А.Б. Летучий