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 Подготовка материалов фонотеки ИРЯ им. В.В. Виноградова РАН для включения в устный корпус в составе Национального корпуса русского языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Касаткина Розалия Францевна, д.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>Москва, 121019, ул. Волхонка, 18/2, (495) 695-26-60
Факс: (495) 695-26-03 roleka@yandex.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования</td><td>ФИО (полностью) руководителя организации – адресата финансирования Молдован Александр Михайлович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Глебова Татьяна Николаевна</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>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="10">Исполнители (ФИО, уч. ст.)</td><td>Каленчук М.Л., д.ф.н., ИРЯ РАН</td></tr>
<tr><td>Савинов Д.М., к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Скачедубова Е.С., к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Щигель Е.В., н.с., ИРЯ РАН</td></tr>
<tr><td>Гришина Е.А., к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Савчук С. О., к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Сердобольская Н.В., к.ф.н., МГППУ</td></tr>
<tr><td>Ловля Е.Н., ИРЯ РАН</td></tr>
<tr><td>Морозова Е.Н., ИРЯ РАН</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 Направление 1. Создание и развитие корпусных ресурсов по современному русскому языку

2. Название проекта Подготовка материалов фонотеки ИРЯ РАН для включения в устный корпус в составе Национального корпуса русского языка

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы) Касаткина Розалия Францевна, д.ф.н., в.н.с
4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)

   Каленчук М.Л., д.ф.н., в.н.с., заместитель директора по научной работе ИРЯ РАН

   Гришина Е. А., к.ф.н., ИРЯ РАН, с.н.с.

   Савчук С. О., к.ф.н., ИРЯ РАН, с.н.с.

   Савинов Д.М., к.ф.н., ИРЯ РАН, с.н.с.

   Скачедубова Е.С., к.ф.н., ИРЯ РАН, н.с.

   Сердобольская Н. В., к.ф.н., МГППУ, преподаватель

   Щигель Е.В., ИРЯ РАН, н.с.

   Ловля Е.Н., ИРЯ РАН, инженер

   Морозова Е.Н., ИРЯ РАН, инженер-программист
5. Содержание фактически проделанной работы, полученные результаты, их новизна, научная и практическая значимость, актуальность, соответствие мировому научному уровню (до 5 стр.)

В 2012 г. работа велась по нескольким направлениям: 1) оцифровка и частичное восстановление записей, находящихся в хранилище Института (прежде всего аналоговых записей на катушках со скоростями 78 и 36 оборо-

тов в минуту); 2) формирование корпуса образцовой литературной речи; 3) пополнение электронной базы данных, описывающей материалы фонотеки.

Результаты работы по первому направлению: были оцифрованы аналоговые записи, сделанных в период 1960-1990-х годов, при этом было обеспечено максимально качественное воспроизведение исходного материала. Общий объем оцифрованного материала около 35 часов звучания.

Результаты работы по второму направлению: были продолжены работы по формированию корпуса образцовой речи носителей русского литературного языка, прежде всего филологов. Отобранные записи отвечают нескольким требованиям: являются содержательно интересными, информационно насыщенными; образцовыми с точки зрения речевого оформления, иллюстрировать не только богатство и разнообразие русской литературной речи, но и индивидуально-авторскую манеру мастера слова. Использованы записи речи разных жанров – научные доклады и выступления (как перед широкой аудиторией, так и в узком кругу), непринужденные беседы, телевизионные и радиопередачи, интервью и пр.

Отобранные записи были качественно улучшены и расшифрованы при помощи специальных программ для анализа и обработки речи PRAAT и Speech Analyzer. В корпус включены расшифровки записей Ю.Д. Апресяна, Т.Г. Винокур, С.С. Высотского, М.Я. Гловинской, М.В. Гординой, В.П. Григорьева, Е.А. Земской, Д.С. Лихачева, М.В. Панова, А.Б. Пеньковского, Н.Д. Светозаровой, А.П. Сковородникова, Ю.С. Степанова и др.

Начато формирование коллекции устной русской речи с региональными особенностями. С этой целью отобраны и расшифрованы записи радиопередач московских и региональных радиостанций: центрального региона России (Радио "Свобода" (Москва), ГТРК "Тула", ГТРК "Липецк"), Севера и Северо-Запада (ГТРК "Архангельск", ГТРК "Карелия", ГТРК "Вятка"), Сибири (Бурятская ГТРК, Радио "Абакан").

Общий объем расшифрованного материала составляет около 30 часов звучания.

Результаты работы по третьему направлению: оцифрованные тексты описаны с помощью системы дескрипторов, задающих внешние характеристики единицы хранения в фонотеке (год записи, сведения о дикторе, жанр и тема текста, время звучания и качество записи), и занесены в базу данных.

Самостоятельным направлением работы над проектом явилась подготовка записей устной речи к размещению на сайте www.ruscorpora.ru/mycorpora-spoken.html . Для этого была произведена дополнительная обработка текстовых расшифровок, которая состояла в следующем.

1. Повторное прослушивание и сверка транскриптов с фонограммой, восстановление и отражение в тексте на основании принятого стандарта особенностей ситуации и произношения диктора (смех, покашливание, шепот и т.д.).

2. Нормализация записей, или сохраняющая разметка, которая состоит в том, что индивидуальным особенностям произношения, которые фиксируются в орфографической записи, ставится в соответствие нормативное написание. Согласно стандартам, принятым в НКРЯ, в транскрипте фиксируются такие особенности устной речи, как стяжки (самые стандартные — *тыща*, *оч*, *щас* и проч.), растяжки (*нууу*, *вооот*) и т.д. Благодаря нормализации записи каждая неправильность сохраняется в тексте, при этом ей приписывается словарная форма, которая, в свою очередь, традиционными для НКРЯ способами, с помощью грамматического парсера, получает свою грамматическую и семантическую разметку. Это, с одной стороны, освобождает от необходимости вводить в автоматический словарь анализатора все неправильные формы, а с другой стороны, сохраняя их в транскрипте, дает возможность изучать именно их в качестве специфических особенностей устной речи.

3. Снабжение текстов социологической разметкой, ее редактирование. Социологическая разметка состоит в приписывании каждому высказыванию сведений о говорящем (пол, возраст или год рождения, профессия), если эти сведения доступны исследователю. В настоящее время социологическая раз-

метка обеспечивает отбор подкорпусов по таким параметрам, как пол говорящего (можно составить корпус мужской или женской речи), возраст говорящего (например, можно составить корпус реплик подростков), год рождения говорящего (можно отобрать реплики дикторов, родившихся в XIX веке), имя диктора.

4. Перевод файлов из текстового формата в формат xml, который производится с помощью специальных программ, завершает подготовку файлов перед размещением их составе корпуса.

5. Метатекстовая разметка текстов, которая состоит в заполнении базы данных, в которой каждому файлу с уникальным именем приписывается набор признаков, характеризующих текст с точки зрения внешних признаков (сведения о дикторе, название текста, место записи и пр.) и собственно текстовых признаков (сфера функционирования, жанр, тип текста и т.д).

6. Снабжение текстов акцентологической разметкой, если предусматривается размещение текстов в составе акцентологического корпуса.

7. Снабжение текстов морфологической и семантической разметкой, которое осуществляется в автоматическом режиме.

В ходе работы над проектом подготовлены и размещены на сайте материалы фонотеки, объем которых составил более 150 тыс. словоупотреблений. Тем самым раздел устной научной речи был пополнен уникальным материалом. Самые ранние из записей относятся к 1958 г.: записи бесед С.С. Высотского с Ю.Н. Андреевой, лекция М.В. Щепкиной в Историческом музее о памятниках письменности на Руси. Остальные материалы относятся к периоду 1960-1990-х годов. Это доклады на конференциях и круглых столах (записи Д.С. Лихачева, Ю.М. Лотмана, В.В. Виноградова, Г.П. Мельникова, Р.И. Аванесова, Л.Л. Касаткина, С.В. Бромлей, И.А. Оссовецкого и др.), выступления на защите диссертации (Т.Г. Винокур, Е.А. Семенец), на заседании ученого совета (А.В. Десницкая, А.П. Евгеньева, В.М. Жирмунский, Ф.П. Сороколетов, Ф.П. Филин) сектора и пр.; выступления деятелей культу-

ры перед слушателями (воспоминания И.Л. Андроникова, Н.О. Гриценко, К.Н. Головко). Кроме того, в устный корпус были включены разножанровые прозаические тексты, среди которых записи непубличной речи, относящиеся к 2007-2008 гг. (беседы в кругу семьи, на работе, ситуативные диалоги, телефонные разговоры), записи учебных и научно-популярных лекций, семинаров, интервью, записи радио- и телепередач, в том числе и религиозного содержания, и пр. Общий объем подготовленного и размещенного материала более 400 тыс. словоупотреблений.

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.)
Grishina, E., Savchuk, S. Multimodal clusters in spoken Russian // LREC 2012 Workshop "Multimodal Corpora: How Should Multimodal Corpora Deal with the Situation?" Istanbul, Turkey, May 21, 2012. ELRA. pp. 10-14.

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)

Оцифрованы аналоговые записи из фондов фонотеки ИРЯ РАН, сделанные в период 1960-1990-х годов, общим объемом 35 часов звучания, при этом обеспечено максимально качественное воспроизведение исходного материала. Продолжено формирование корпуса образцовой речи носителей русского литературного языка, прежде всего филологов. Включены записи речи раз-

ных жанров – научные доклады и выступления (как перед широкой аудиторией, так и в узком кругу), непринужденные беседы, интервью и пр. Начато формирование коллекции устной русской речи с региональными особенностями: отобраны и расшифрованы записи радиопередач московских и региональных радиостанций. Общий объем расшифрованных записей - 30 часов звучания. Пополнен электронный каталог документов фонотеки, содержащий описание единиц хранения в фонотеке с помощью системы дескрипторов (год записи, сведения о дикторе, жанр и тема текста, время звучания и качество записи). Материалы фонотеки, объемом около 150 тыс. словоупотреблений, прошли специальную обработку и размещены на сайте НКРЯ. Кроме того, в состав устного подкорпуса НКРЯ были включены разножанровые прозаические тексты, среди которых записи непубличной речи, относящиеся к 2007-2008 гг. (беседы в кругу семьи, на работе, ситуативные диалог), записи учебных и научно-популярных лекций, радио- и телепередач, и пр. Общий объем подготовленного и размещенного материала более 400 тысяч словоупотреблений.

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

Пополнен электронный архив фонотеки ИРЯ РАН новыми оцифрованными аудиозаписями в объеме 35 часов звучания. Расшифрованы материалы в объеме 30 часов звучания. Подготовлены к размещению в устном корпусе НКРЯ тексты разных типов в объеме 400 тыс. словоупотреблений, в том числе уникальные записи из фондов фонотеки ИРЯ РАН.

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 году планируется пополнение электронного архива фонотеки ИРЯ РАН оцифрованными записями, представляющими образцы литературной русской речи с региональными особенностями в объеме 35 часов звучания. Будет продолжено формирование корпуса образцовой литературной речи, который пополнится расшифровками материалов фонотеки в объеме 30 часов звучания. Для устного модуля НКРЯ будут подготовлены тексты разных типов в объеме не менее 300 тыс. словоупотреблений.

Подпись руководителя проекта

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
	Подготовка материалов фонотеки ИРЯ РАН для включения в устный корпус в составе Национального кор-пуса русского языка	ИРЯ РАН	Р.Ф. Касаткина		2013 г. – Пополнение электронного архива фоно-теки ИРЯ РАН оцифрованными записями, пред-ставляющими образцы литературной русской ре-чи с региональными особенностями в объеме 35 часов звучания. Расшифровка материалов фоно-теки в объеме 30 часов звучания. Подготовка к размещению в устном корпусе текстов разных типов в объеме 300 тыс. словоупотреблений.