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 Развитие мультимедийного модуля Национального корпуса русского языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Гришина Елена Александровна, к.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>rudi2007@yandex.ru</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>Савчук С. О., к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Иванютин С.Б., РГГУ</td></tr>
<tr><td>Курсакова А.А., РГГУ, аспирант</td></tr>
<tr><td></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 **Направление 1. Создание и развитие корпусных ресурсов по современному русскому языку**

2. Название проекта **Развитие мультимедийного модуля Национального корпуса русского языка**

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы) **Гришина Елена Александровна, к.ф.н., с.н.с., Институт русского языка РАН**

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

Савчук Светлана Олеговна, к.ф.н., ИРЯ РАН, с.н.с.
Курсакова А.А., РГГУ, аспирант
Иванютин С.Б., РГГУ, аспирант

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 все задачи, поставленные на этот год, были выполнены.

1. Пополнение Мультимедийного русского корпуса (МУРКО).

1.1. Зона «Речь кино». В результате осуществления проекта в 2012 году Мультимедийный русский корпус, в пилотном варианте содержавший 75 фильмов, был существенно расширен и теперь может считаться не пробным (пилотным), а полноценно функционирующим мультимедийным корпусом, позволяющим решать самые разные задачи, связанные с исследованием русской устной речи.

В ходе работы по Программе РАН 2012-2014 гг. в МУРКО в течение отчетного года были добавлены 109 фильмов (список см. в Приложении 1), или 28 028 клипов, общим объемом порядка 100 часов звучания. С учетом результатов Программы РАН «Корпусная лингвистика» 2011 года, а также гранта РФФИ 10-06-00151а «Разработка и создание Мультимедийного

русского корпуса (МУРКО) в рамках Национального корпуса русского языка (www.ruscorpora.ru)», на конец 2012 года объем зоны «Речь кино» в Мультимедийном русском корпусе составляет 439 фильмов, более 3 млн словоупотреблений.

1.2. Публичная/непубличная речь. В 2012 году была продолжена работа по пополнению некинематографической зоны МУРКО, прежде всего, – зоны публичной речи. В корпусе были размещены тексты следующих жанров:

- рассказ
- воспоминания
- беседа
- научный доклад
- дискуссия
- радиореклама

Общий объем – более 40 тыс. словоупотреблений, общая длительность звучания – более 3 часов.

Огромный интерес представляют включенные в МУРКО документальные фильмы, созданные на телеканале «Культура» в серии «Письма из провинции». Они представляют собой монологи перед камерой жителей самых разных регионов нашей страны, разных профессий, которые рассказывают о своей профессиональной деятельности, о своем творчестве, о повседневной жизни провинциальной России. За отчетный период в корпус включены следующие фильмы, общим объемом более 20000 словоупотреблений, длительность звучания около 1,5 часов.

- Где-то там в Омутнинске.
- Станица Тамань. Таманские казаки.
- Уютное гнездо. Старая Русса.
- В холмогорских снегах. Холмогоры.
- Старица. Выход.
- Иркутск.

1.3. Авторское/художественное чтение. Эта зона является новой для Мультимедийного русского корпуса. Необходимость ее создания определяется тем, что если в отношении синтаксиса, словообразования, лексики тексты, включенные в эту часть корпуса, не имеют особых отличий от корпуса письменных текстов, то в отношении фонетики, интонации, ударения, коммуникативной структуры (функционирование дискурсивных маркеров, порядок слов, типы и композиция пауз, типы эмфатических выделений) тексты типа written to be spoken имеют огромную ценность. В ходе выполнения программы в эту зону МУРКО были включены тексты М.М. Пришвина («Мои тетрадки», «Кукушка»), М.М. Зощенко («Расписка»), В.Т. Шаламова («Белка», «Воскрешение лиственницы») – все в авторском исполнении; рассказы А.П. Чехова в актерском исполнении, в том числе театрализованном. Общее время звучания – примерно 1 час.

2. Аннотация корпуса.

Все подготовленные для МУРКО тексты были аннотированы по стандартной для мультимедийного подкорпуса НКРЯ методике. Разметка делится на три группы:

2.1. Стандартная разметка Национального корпуса русского языка включает в себя метатекстовую аннотацию (т.е. характеристику текста с точки зрения автора, даты создания, жанра, тематики и др.), морфологическую и семантическую аннотацию.

2.2. Разметка устного модуля НКРЯ включает в себя акцентологическую разметку (постановка реального ударения в словоформе или указание на отсутствие ударения) и социологическую разметку (указание на пол, год рождения, возраст и имя говорящего, которое приводится при каждой реплике).

2.3. Специфическая разметка МУРКО, которая включает в себя орфоэпическую аннотацию (т.е. указание на сочетания гласных и согласных внутри слова и на границах слов) и разметку вокалической структуры слова (разметка номера и качества ударного гласного, номера и качества предударного или заударного гласного, количество слогов в словоформе).

Все типы разметки могут комбинироваться, что дает возможность пользователю получать в результате грамотно построенного запроса корпусную выдачу с минимальным количеством шума.

3. Мультимедийный параллельный корпус. Подготовительные работы

В ходе выполнения Программы РАН 2011 г. нами в общих чертах была разработана архитектоника Мультимедийного параллельного корпуса (МультиПАРКа), который предназначен для сопоставительных внутри- и межъязыковых исследований устной речи. Вкратце концепция этого корпуса была изложена в двух статьях:

а) *Гришина Е. А.* Мультимедийный русский корпус (МУРКО): современное состояние и перспективы развития // Труды международной конференции «Корпусная лингвистика – 2011», СПб., 27-30 июня 2011, с. 138-144

б) *E. Grishina, S. Davydov, S. Savchuk, D. Sichinava, A. Zobnin.* Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure // The 8th international conference on Language Resources and Evaluation (LREC) – 2012 http://www.corpora.uni-hamburg.de/lrec2012/Proceedings_Complete.pdf

В течение 2012 года

3.1. Была дополнительно разработана концепция **англо-русского параллельного видеокорпуса** (МультиАРК), который должен включать в себя, по предварительным планам, американские и английские фильмы (сериалы), нарезанные на клипы (по методологии, отработанной в ходе реализации Мультимедийного русского корпуса), выровненные с соответствующими расшифровками звуковой дорожки на английском языке, а также с соответствующими им участками звуковой дорожки русского перевода. При этом предполагается, что английский текст оригинала и русский текст дубляжа будут морфологически размечены так, как английский и русский текст раз-

мечаются в Параллельном англо-русском подкорпусе Национального корпуса русского языка. МультиАРК позволит:

- проводить исследования английской устной речи
- проводить сопоставительные исследования английской и русской устной речи (привлекая для этого данные из основной зоны МУРКО) с точки зрения
    - жестикуляции
    - фонетики
    - паузации
    - интонирования

3.2. Был подобран материал для **пилотного корпуса МультиПАРК**, который, как предполагается, будет включать в себя пьесу Гоголя «Ревизор» в следующих реализациях:

- Аудиоспектакль «Ревизор» 2006 г.
- Аудиокнига «Ревизор» в исполнении А. Клюквина
- Аудиокнига «Ревизор» в исполнении С. Юрского
- Радиоспектакль «Ревизор», 1949 г.
- Спектакль «Ревизор» театра им. Л. Украинки, Киев, 2002 г.
- Спектакль «Ревизор» Александринского театра, СПб., 2003
- Спектакль «Ревизор» театра Сатиры, Москва
- Телеспектакль «Ревизор» Малого театра, Москва
- Кинофильм «Ревизор» С. Газарова, 1996 г.
- Телеспектакль «Ревизор. Сцены» Большого драматического театра, Ленинград, 1972
- Кинофильм «Ревизор» В. Петрова, 1952 г.
- Кинофильм «Инкогнито из Петебурга» Л. Гайдая, 1977
- Спектакль «Ревизор» Театра-студии О. Табакова, постановка С. Газарова

Таким образом, в ходе выполнения программы в 2012 году был создан и активно функционирует Мультимедийный корпус русского языка (МУРКО), который по объему и разнообразию включенного в него материала, а также по богатству разметки не имеет аналогов среди открытых мультимедийных корпусов. Он предоставляет уникальную возможность исследовать устную русскую речь с точки зрения фонетики, интонации, синтаксиса, жестикуляции, стилистики, коммуникативных стратегий, структуры и типологии жанров устной речи и др. Мы предполагаем и дальше развивать этот уникальный ресурс.

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

6.1. количество монографий –

6.2. количество сборников статей –

6.3. количество статей 5

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.)

1. Автодейксис: основные типы и значения // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2012», с. 173-186
2. *Здесь* и *тут*: корпусной и жестикуляционный анализ полных синонимов // "Русский язык в научном освещении", № 23, 2012, с. 39-71
3. Указания рукой как система (по данным Мультимедийного русского корпуса) // «Вопросы языкознания», № 3, 2012, с. 3-50

9. Список книг, сданных в печать или поданных на издательские гранты (указать ожидаемое время издания, объем в а.л.) 1 (2013, 10 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)

Объем Мультимедийного русского корпуса (МУРКО) на конец 2012 года составляет более 3 млн словоупотреблений. Он включает в себя 439 художественных фильмов периода 1930–2008 гг., а также зону некинематографической речи объемом более 100 тыс. словоупотреблений (публичная и непубличная устная речь, авторское и художественное чтение, театральная речь). Общий объем звучания в корпусе превосходит 400 часов. Разработана концепция Мультимедийного параллельного корпуса (МультиПАРК), собран материал, предназначенный для включения в этот корпус. Также разработана общая архитектоника англо-русского параллельного видеокорпуса (МультиАРК). На материале МУРКО написаны и опубликованы научные статьи. Таким образом, в ходе выполнения программы в 2012 году был создан и активно функционирует Мультимедийный корпус русского языка (МУРКО), который по объему и разнообразию включенного в него материала, а также по богатству разметки не имеет аналогов среди открытых мультимедийных корпусов в мире. Он предоставляет уникальную возможность исследовать устную русскую речь с точки зрения фонетики, интонации, синтаксиса, жестикуляции, стилистики, коммуникативных стратегий, структуры и типологии жанров устной речи и др. Мы предполагаем и дальше развивать этот уникальный ресурс, ориентируя его как на новые жанры, так и на сопоставительные исследования.

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

Объем Мультимедийного русского корпуса (МУРКО) на конец 2012 года составляет более 3 млн словоупотреблений. Он включает в себя 439 художественных фильмов периода 1930–2008 гг., а также зону некинематографической речи объемом более 100 тыс. словоупотреблений. Общий объем звучания в корпусе превосходит 400 часов. Разработана концепция Мультимедийного параллельного корпуса (МультиПАРК), собран материал, предназначенный для включения в этот корпус.

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) Пополнение подкорпуса «Речь кино» в составе МУРКО: количество фильмов в подкорпусе будет доведено до 530, что в итоге даст нам мультимедийный корпус объемом порядка 3,5 млн словоупотреблений или около 400 часов звучания.
2) Пополнение подкорпуса публичной речи (в общей сложности примерно на 0,5 млн словоупотреблений, или 50 часов звучания) за счет целенаправленного формирования подкорпуса научного и политического дискурса.
3) Формирование подкорпуса театральной речи, что в результате позволит изучать соотношение «естественной» публичной речи, речи театра и речи кино с опорой на солидные статистические и содержательные данные (предполагаемый объем – порядка 50 тыс. словоупотреблений, или 5 часов звучания)
4) Разработка программного обеспечения (программа Scripter), которая оптимизирует процесс обработки текстов, входящих в состав МУРКО, а также позволит максимально быстро упорядочивать тексты, подготовленные для Мультимедийного параллельного корпуса (МультиПАРК).
5) Подготовка и обработка расшифровок, видео- и аудиодорожек для корпуса МультиПАРК (различные сценические и кинематографические воплощения пьесы Гоголя «Ревизор»).
6) Подготовка материалов для пилотного англо-русского параллельного видеокорпуса в составе МультиПАРКа (предварительно название этого корпуса – МультиАРК, Мультимедийный англо-русский корпус).

Подпись руководителя проекта Е.А. Гришина

Приложение 1. Список фильмов, добавленных в МУРКО в 2012 г.

	Название	Количество клипов
1.	Неисправимый лгун	203
2.	Необыкновенный концерт (кукольный спектакль)	189
3.	Неоконченная повесть	196
4.	Неоконченная пьеса для механического пианино	313
5.	Непобедимые	228
6.	Неподдающиеся	213
7.	Неподсуден	205
8.	Неуловимые мстители	139
9.	Новая Москва	207
10.	Новые приключения неуловимых	209
11.	Новый Гулливер	100
12.	Ночной дозор	305
13.	Ночной продавец	175
14.	О бедном гусаре замолвите слово	596
15.	Облако-рай	212
16.	Овод	256
17.	Одиноким предоставляется общежитие	209
18.	Одиночное плавание	210
19.	Однажды летом	119
20.	Они сражались за родину	346
21.	Опасно для жизни	232
22.	Опасные гастроли	194
23.	Операция Ы и другие приключения Шурика	155
24.	Оптимистическая трагедия	264
25.	Осенний марафон	257
26.	Остров	227
27.	Отец солдата	198
28.	Отпуск в сентябре	370
29.	Отроки во вселенной	221
30.	Отряд особого назначения	112
31.	Офицеры	242
32.	Ошибка инженера Кочина	294
33.	Павел Корчагин	267
34.	Папа	237
35.	Парад планет	136
36.	Парень из нашего города	211
37.	Партийный билет	219
38.	Пацаны	236
39.	Первая перчатка	299
40.	Переходный возраст	245
41.	Петр Первый	481
42.	Петя по дороге в Царствие Небесное	178
43.	Печки-лавочки	261
44.	Пираты XX века	126
45.	Письма мертвого человека	153
46.	ПитерFM	205
47.	Плюмбум, или Опасная игра	221
48.	По главной улице с оркестром	244
49.	По семейным обстоятельствам	359
50.	Повесть о настоящем человеке	179
51.	Повесть о первой любви	168
52.	Подвиг разведчика	261
53.	Подводная лодка Т-9	117
54.	Подранки	206
55.	Поединок	183
56.	Покровские ворота	402
57.	Полеты во сне и наяву	171

	Название	Количество клипов
58.	Полосатый рейс	227
59.	Попса	283
60.	Послесловие	308
61.	Почти смешная история	497
62.	Праздник Нептуна	130
63.	Праздник святого Йоргена	43
64.	Приваловские миллионы	523
65.	Приключения Буратино	304
66.	Приключения капитана Врунгеля, м/ф	177
67.	Приключения Шерлока Холмса и доктора Ватсона. Двадцатый век начинается	433
68.	Приключения Шерлока Холмса и доктора Ватсона. Собака Баскервилей	386
69.	Приключения Шерлока Холмса и доктора Ватсона. Сокровища Агры	373
70.	Принцесса на горошине	194
71.	Приходите завтра…	257
72.	Про Красную Шапочку	401
73.	Про уродов и людей	130
74.	Проверка на дорогах	200
75.	Пропавшая экспедиция	340
76.	Простая история	247
77.	Простые вещи	196
78.	Прохиндиада, или Бег на месте	246
79.	Путешествие с домашними животными	135
80.	Путь в «Сатурн»	217
81.	Пятнадцатилетний капитан	363
82.	Пять вечеров	292
83.	Раба любви	248
84.	Ребро Адама	211
85.	Республика ШКИД	302
86.	Родня	219
87.	Розыгрыш	322
88.	Свадьба (реж. П. Лунгин)	273
89.	Свадьба в Малиновке	223
90.	Свадьба с приданым	308
91.	Свадьба	179
92.	Сватовство гусара	132
93.	Светлый путь	227
94.	Свинарка и пастух	175
95.	Свой среди чужих, чужой среди своих	185
96.	Свои	211
97.	Связь	183
98.	Сдается квартира с ребенком	205
99.	Сегодня – новый аттракцион	261
100.	Секретная миссия	238
101.	Семеро смелых	199
102.	Семнадцать мгновений весны	1855
103.	Семь нянек	206
104.	Семь стариков и одна девушка	214
105.	Сердца четырех	253
106.	Сережа	207
107.	Сибирский цирюльник	459
108.	Сказание о земле сибирской	215
109.	Сказка странствий	255
ИТОГО: 28 028 клипов, более 100 часов звучания		

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
	Развитие мультимедийного модуля Национального корпуса русского языка	ИРЯ РАН	Е.А.Гришина + 5 человек		1) Пополнение подкорпуса «Речь кино» в составе МУРКО: количество фильмов в подкорпусе будет доведено до 530, что в итоге даст нам мультимедийный корпус объемом порядка 4 млн словоупотреблений или около 400 часов звучания. (200 тыс. руб.) 2) Пополнение подкорпуса публичной речи (в общей сложности примерно на 0,5 млн словоупотреблений, или 50 часов звучания) за счет целенаправленного формирования подкорпуса научного и политического дискурса. (100 тыс. руб.) 3) Формирование подкорпуса театральной речи, что в результате позволит изучать соотношение «естественной» публичной речи, речи театра и речи кино с опорой на солидные статистические и содержательные данные (предполагаемый объем – порядка 50 тыс. словоупотреблений, или 5 часов звучания) (100 тыс. руб.) 4) Разработка программного обеспечения (программа Scripter), которая оптимизирует процесс обработки текстов, входящих в состав МУРКО, а также позволит максимально быстро упорядочивать тексты, подготовленные для Мультимедийного параллельного корпуса (МультиПАРК). (100 тыс. руб.) 5) Подготовка и обработка расшифровок, видео- и аудиодорожек для корпуса МультиПАРК (различные сценические и кинематографические воплощения пьесы Гоголя «Ревизор»). (100 тыс. руб.) 6) Подготовка материалов для пилотного англо-русского параллельного видеокорпуса в составе МультиПАРКа (предварительно название этого корпуса – МультиАРК, Мультимедийный англо-русский корпус). (100 тыс. руб.)