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 Пополнение и развитие акцентологического корпуса русского языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.) Каленчук Мария Леонидовна, д.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>mkalenchuk@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>Телефон, факс (с кодом города), E-mail главного бухгалтера организации – адресата финансирования (495) 695-26-60 Факс: (495) 695-26-03</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.
500000</td></tr>
<tr><td rowspan="6">Исполнители (ФИО, уч. ст.)</td><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>Шестакова Л.Л., к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Корчагин К.М., ИРЯ РАН</td></tr>
<tr><td>Поляков А.Е., НПБ им. К.Д. Ушинского</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 Направление 1. Создание и развитие корпусных ресурсов по современному русскому языку

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы) Каленчук Мария Леонидовна, д.ф.н., в.н.с., заместитель директора по научной работе

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

   Гришина Елена Александровна, к.ф.н., ИРЯ РАН, с.н.с.

   Савчук Светлана Олеговна, к.ф.н., ИРЯ РАН, в.н.с.

   Сердобольская Наталья Вадимовна, к.ф.н., РГГУ, преподаватель

   Шестакова Лариса Леонидовна, к.ф.н., ИРЯ РАН, в.н.с.

   Корчагин Кирилл Михайлович, ИРЯ РАН, аспирант

   Поляков Алексей Евгеньевич, ФГУП НПБ им. К.Д. Ушинского, с.н.с.

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

Акцентологический корпус, созданный в составе Национального корпуса русского языка, предоставляет возможность изучать словесное ударение не на основе словарей, а наблюдая реальные тексты. В составе корпуса выделяются две зоны. Зона прозаических текстов отражает современное состояние ударения и включает записи устной речи, в которых ударение расставле-

но в соответствии с реальным произношением. Тексты относятся к разным функциональным сферам: транскрипты кинофильмов, образцы спонтанной бытовой речи, публичной устной речи разной степени спонтанности.

Поэтическая зона содержит тексты, в которых размечены слоги, на которые может падать ударение. Эта зона в настоящий момент в основном отражает историю русского ударения, поскольку в корпусе представлена поэзия XVIII – начала XX вв. Эти тексты служат в русистике традиционным источником для исследования норм словесного ударения предшествующих эпох.

В ходе осуществления проекта в 2012 году для пополненеия Акцентологического корпуса был подготовлен новый материал.

1. Пополнение зоны прозаических текстов.

1.1. Пополнение собственно устной зоны (публичная и непубличная естественная устная русская речь). В эту зону включены:

– записи повседневной (бытовой) устной речи: беседы со знакомыми, в семейном кругу, рассказы-воспоминания, ситуативные диалоги в магазине, на рынке, на вокзале и т.д., которые отражают спонтанную речь носителей как старшей, так и младшей нормы;

– документальные фильмы, теле- и радиопередачи, записанные и снятые в разных городах, что значительно расширило географию корпуса. В нем появились расшифровки записей, сделанных в городах и поселках Центральной России (Калужская обл., Ивановская обл., Нижегородская обл., Волгоградская обл., Воронеж), на Севере и Северо-западе (Архангельская обл., Вологодская обл., Кировская обл., Ленинградская обл., Калининградская обл., Псковская обл., Карелия, Ненецкий АО), на юге России (Краснодарский край), в Сибири (Кемерово, Кузбасс, Орск).

– лекции, доклады, диспуты и прочие публичные выступления. В корпус включены материалы фонотеки ИРЯ РАН - расшифровки записей известных филологов Р.И. Аванесова, В.В. Виноградова, Т.Г. Винокур,

С.С. Высотского, Д.С. Лихачева, М.Ю. Лотмана, Г.П. Мельникова, М.В. Панова, А.Б. Пеньковского, Ф.П. Филина, и др., относящиеся к 1960-1980 гг.

– записи бесед и выступлений перед слушателями деятелей культуры (И.Л. Андроников, Н.О. Гриценко, К.Н. Головко);

– церковные проповеди, беседы (например, беседы протоиерея Олега Стеняева, священника Андрея Лоргуса, проповеди о. Георгия Чистякова);

– записи открытых судебных заседаний, фиксирующие речь участников судебных прений и отражающие новый для корпуса тип судебного дискурса.

1.2. Продолжена разработка и наполнение контентом раздела «Художественное чтение», который включает в себя прозаические тексты в исполнении их авторов и чтецов-актеров. Необходимость его 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 тексты типа written-to-be-spoken имеют огромную ценность. Эта зона является новой для Акцентологического корпуса и до сих пор содержала образцы авторского чтения. В ходе выполнения проекта в 2012 году в эту зону были включены произведения в актерском исполнении: рассказы А.П. Чехова в исполнении актеров старшего поколения О. Абдулова, Р. Плятта, Ф.Г. Раневской, Н. Якушенко и т. д.

2. **Подготовка текстов к размещению на сайте**

Подготовка к размещению на сайте www.ruscorpora.ru/mycorpora-accent.html включает расшифровку аудиозаписей, повторное прослушивание и сверку транскриптов с фонограммой, акцентуирование текстов программными средствами и последующее редактирование акцентологической разметки, снабжение текстов метатекстовой, морфологической и семантической разметкой.

В Акцентологическом корпусе используются те же 4 вида разметки, что и в устном корпусе, но с некоторыми особенностями. Например, в *метатексто-*

вой аннотации используются не только признаки, общие для зон поэзии и прозы (имя автора, возраст, пол, дата записи текста), но также параметры, специфические для каждой зоны. Для поэтических текстов это метр, клаузула, рифма, размер, для прозы – тип текста.

Морфологической и *семантической* разметкой снабжена каждая словоформа, а *социологическая* аннотация, с помощью которой каждому высказыванию приписывается информация о говорящем, используется только для зоны прозы.

Кроме того, используется специфическая акцентологическая аннотация, благодаря которой каждая словоформа получает знак ударения (или не получает в случае безударности), что позволяет строить запросы и получать сведения об ударных и безударных словоформах в комбинации с грамматическими и семантическими признаками. Акцентологическая аннотация осуществлялась в два этапа.

Первый этап выполнялся в автоматическом режиме с помощью программы Accentuator (автор А. Поляков), которая расставила ударения в словах обрабатываемого транскрипта в соответствии с рекомендациями встроенного в нее словаря. Этот словарь построен на базе нормативных словарей русского языка, но существенно дополнен разработчиками корпуса.

На втором этапе лингвист-эксперт прослушивал аудиозаписи и вносил поправки в размеченный текст. В результате получался текст, в котором ударения расставлены в соответствии с реальным произношением.

В 2012 году разработана и освоена специальная программа-редактор для проверки правильности акцентологической аннотации. В ней используется система одного окна для прослушивания аудиофайла и редактирования транскрипта, а также предусмотрено выделение потенциально ошибочных разборов, что значительно облегчает процесс редактирования.

Последующая обработка заключалась в переводе файлов из текстового формата в формат xml, который производится с помощью последовательного

применения специальных программ, и завершает подготовку файлов перед размещением их составе корпуса.

Снабжение текстов метатекстовой разметкой состояло в заполнении базы данных, в которой каждому файлу с уникальным именем приписывался набор атрибутов, характеризующих текст с точки зрения внешних признаков (сведения о дикторе, название текста, место записи и пр.) и собственно текстовых категорий (сфера функционирования, жанр, тип текста и т.д).

Общий объем нового материала, подготовленный для пополнения прозаической зоны Акцентологического корпуса, составил более 0,4 млн словоупотреблений.

3. Пополнение зоны поэтических текстов.

Поэтическая зона содержит тексты, в которых размечены слоги, на которые может падать ударение. Путем пересчета размеченных сильных долей по определенным правилам мы можем получить точные сведения о месте ударения в том или ином слове. Эта зона в настоящий момент в отражает не только историю русского ударения (поэзия XVIII – начала XX вв.), но и его современное состояние, поскольку в поэтическую зону входят тексты второй половины XX в. В рамках проекта с акцентологической точки зрения размечены и включены в корпус поэтические тексты XX в. в объеме 1 млн словоупотреблений.

Таким образом, в настоящее время акцентологический корпус содержит более 12 млн словоупотреблений. Тексты распределены по двум зонам следующим образом:

Зона		**Объем, млн словоупотр.**	**Доля**
Поэзия		7,7	61,5%
Проза	речь кино	3,8	30,2%
	публичная речь	0,9	7,1%
	непубличная речь	0,1	0,8%
	чтение	0,07	0,4%

Примерное распределение по периодам:

Период	Поэзия, млн словоупотреблений	Проза, млн словоупотреблений
1700 -1799	1	
1800-1910	4,2	0,004
1911-1949	1,9	0,4
1950-1979	0,4	2,1
1980-1999	0,006	1
2000-2008	-	1,2

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)

Пополнение акцентологического корпуса новыми текстами, общим объемом около 1,4 млн словоупотреблений. Из них поэтических текстов – около 1 млн словоупотреблений, прозаических текстов – 0,4 млн словоупотреблений. В новый раздел «Художественное чтение», наряду с авторским чтением включены произведения в актерском исполнении. Разработано программное обеспечение, повышающее качество редактирования транскриптов, произведена коррекция акцентологической разметки. Прозаические тексты представляют

устную речь в функциональных, территориальных, возрастных разновидностях, что дает материал для изучения варьирования акцентных норм.

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

Пополнение акцентологического корпуса новыми текстами, общим объемом около 1,4 млн словоупотреблений. Из них поэтических текстов – около 1 млн словоупотреблений, прозаических текстов – 0,4 млн словоупотреблений. В новый раздел «Художественное чтение», наряду с авторским чтением включены произведения в актерском исполнении. Разработано программное обеспечение, повышающее качество редактирования транскриптов, произведена коррекция акцентологической разметки.

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 пополнение корпуса новыми текстами, общим объемом около **1,5 млн** словоупотреблений. Прозаических текстов – около 0,5 млн словоупотреблений, включая записи театральных постановок, поэтических текстов – 1 млн словоупотреблений. Предполагается совершенствование социологической аннотации.

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
	Пополнение и развитие акцентологического корпуса русского языка	ИРЯ РАН	М.Л. Каленчук		2013 г. - Пополнение корпуса новыми текстами, общим объемом около **1,5 млн** словоупотреблений. Прозаических текстов – около 0,5 млн словоупотреблений, включая запи-си театральных постановок, поэтических текстов – 1 млн словоупотреблений. Совершенствование социологической аннотации