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 Системное пополнение и совершенствование организации корпусов НКРЯ с базовой разметкой (диахронического, газетного)</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Шайкевич Анатолий Янович, д.ф.н</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>119019 Москва, ул. Волхонка 18/2 ИРЯ РАН
irlras@mail.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования

Федеральное государственное бюджетное учреждение науки 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>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Глебова Татьяна Николаевна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
119019 Москва, ул. Волхонка 18/2
телефон: (495) 695 28 07</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="3">Исполнители (ФИО, уч. ст.)</td><td>Савчук С.О., кфн, ИРЯ РАН</td></tr>
<tr><td>Сичинава Д.В., кфн, ИРЯ РАН</td></tr>
<tr><td>Гришина Е.А., кфн, ИРЯ РАН</td></tr>
<tr><td></td><td>Ловля Е.Н., ИРЯ РАН</td></tr>
<tr><td></td><td>Морозова Е.Н., ИРЯ РАН</td></tr>
<tr><td></td><td>Смирнова Е.А., 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 Системное пополнение и совершенствование организации корпусов НКРЯ с базовой разметкой (диахронического, газетного)

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)
Шайкевич Анатолий Янович, д.ф.н., г.н.с. Института русского языка им. В.В. Виноградова РАН

4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)
Гришина Е.А., к.ф.н, с.н.с. ИРЯ РАН, Летучий А.Б. кфн, ИРЯ РАН, Савчук С.О., кфн, ИРЯ РАН, Сичинава Д.В., кфн, ИРЯ РАН, Смирнова Е.А., мнс, ИРЯ РАН, Ловля Е.Н., ИРЯ РАН, инженер, Морозова Е.Н., ИРЯ РАН, инженер

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

Проект направлен на развитие корпуса письменных текстов 1-й половины XX в. - одного из текстовых массивов в составе основного корпуса НКРЯ. Собранные в нем тексты представляют огромный интерес для исследователей в таких областях, как история русского литературного языка и ее периодизация, особенности языка советской эпохи, язык русского зарубежья, историческое изучение языковых явлений в их соотношении с процессами в социокультурной жизни общества и др. Созданный в 2006-2008 гг. корпус в настоящее время нуждается в лингвистической и программной поддержке: в пополнении состава текстов, совершенствовании лингвистической разметки

и архитектуры, коррекции ошибок, расширении функциональных возможностей поиска, сохранения и обработки данных.

В ходе осуществления проекта в 2012 г. были выполнены следующие виды работ.

1. Анализ текстового состава корпуса по социологическим и жанрово-стилистическим параметрам.

Предварительный анализ существующего баланса текстов корпуса необходим по той причине, что его результатами будет определяться отбор текстов по заданным критериям и их определенное соотношение при пополнении уже функционирующего корпуса. Анализ текстового состава корпуса по состоянию на 2012 год выявил следующую картину. Распределение текстов основного корпуса по сферам коммуникации представлено в таблице 1.

Таблица 1

сфера	**1-я пол. XX в.**	**%**	**НКРЯ в целом**	**%**
всего с/у	54 021 304		209 203 107	
художественная	25861162	47%	92 060 453	44%
публицистика	16697987	30%	76 002 916	36%
учебно-научная	8541284	16%	27 375 161	13%
официльно-деловая	1124611	2%	3 854 693	1,8%
бытовая	1466148	2,7%	4 443 085	2%
церк-богословская	596542	1,1%	3 222 162	1,5%
произв-техническая	381759	1%	1 326 442	0,6%
реклама	14948	>0,1%	572 414	0,3%
электр. коммуникация	0	0	2 286 557	1%

Как видно из таблицы, с точки зрения распределения текстов по разным сферам функционирования корпус 1-ой пол. XX в. является весьма сбалансированным. Поэтому при включении новых текстов следует сохранять существующие пропорции и в равной степени уделять внимание текстам всех функциональных сфер.

Анализ текстового состава в хронологическом отношении выявил неравномерное распределение текстов по десятилетиям (см. рис. 1)

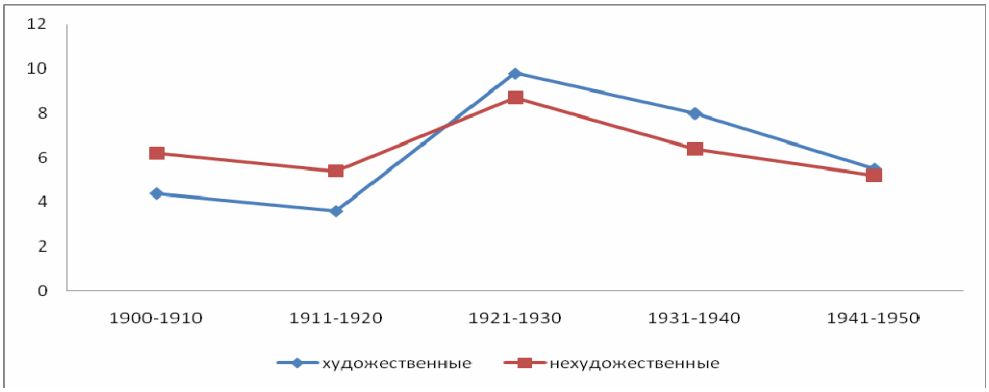

Рис. 1.

Как видно из диаграммы, в корпусе обнаружился недостаток текстов, относящихся к периоду 1900-1920 гг., а также 1941-1950 гг. Поэтому вырав-

нивание корпуса в хронологическом отношении предполагает приоритетное включение в него текстов данных периодов.

Анализ текстов в тематическом отношении выявил преобладание в корпусе 1-ой пол. XX в. текстов, относящихся к гуманитарным сферам (философия, история, филология, культуролоия, искусство и культура, здоровье и медицина) – до 70%. Эту диспропорцию следует учесть при включении новых текстов научного содержания.

Таким образом, практическим результатом проведенного анализа стало выявление тех участков речевого употребления, которые представлены в корпусе недостаточным количеством текстов, на основании чего был разработан план дальнейшего расширения.

2. Пополнение корпуса первой половины XX века текстами.

В 2012 году в корпус были значительно пополнен текстами, относящимися к тем сферам функционирования и к тем периодам, которые недостаточно в нем представлены. Следует отметить наиболее представительные коллекции текстов:

- газетно-журнальная периодика: российские газеты за 1911 и 1912 год, советские газеты и журналы «Советское искусство», «Нижегородский кооператор», «Советская архитектура», «Смена», «Русский спорт» и др. за 1919, 1926, 1934, 1939, 1942,1945, 1950-1952 гг.; эмигрантские журналы «Новый дом» (1926-1927 гг.), «Новый корабль» (1927 г.)
- мемуары: воспоминания современников о Чехове, Горьком, Маяковском, Есенине, Суворине, воспоминания Т. В. Солоневич, П.А. Моисеенко, К.С. Петров-Водкина и др.
- публицистика и литературная полемика: статьи и выступления представителей различных политических партий – эсеров (В.М. Чернов, Б.В. Савинков, М.В. Вишняк, В.М. Зензинов), кадетов (П.Н. Милюков, В.А. Маклаков, А.А. Корнилов, В.Д. Набоков), протоколы совещания членов Учредительного собрания 1921 г.; манифесты представителей различных литературных направлений начала XX в. - символистов, акмеистов, имажинистов, футуристов, конструктивистов, ОБЭРИУтов, РАПП и др.
- научно-популярная литература: В.К. Арсеньев, Я.И. Перельман и др.
- художественная литература: С.И. Гусев-Оренбургский, К.А. Коровин, А.С. Новиков-Прибой, С.Н. Сергеев-Ценский, А. Сорокин, Л.А. Чарская и др.).

Общий объем подготовленного и размещенного на сайте материала - более 4 млн. словоупотреблений.

3. Подготовка электронных версий текстов.

Подготовка значительной части текстов требовала проведения полного цикла работ, включая сканирование и распознавание текстов из форматов .pdf, .tif, .jpg, .djvu, поскольку многие тексты (газеты, документы и пр.) представлены в электронных библиотеках в графических форматах.

Редактирование электронных версий текстов дореволюционных изданий, связанное с орфографической модернизацией, осуществлялось в соответствии с эдиционными принципами, принятыми для изданий академического типа или близких к ним, в том числе для филологических электронных библиотек. При воспроизведении текстов, изданных до 1956 года, а также дореволюционных и эмигрантских изданий в них сохранялись все особенности орфографических норм соответствующего периода, за исключением тех изменений в графике, которые были внесены реформой 1918 года. Отредактированные версии текстов в дореволюционной графике сохраняются в архиве.

4. Коррекция ошибок и совершенствование морфологической аннотации.

Множественность орфографических вариантов передачи одного и того же слова или формы, представляющая интерес для специалистов, изучающих историю и современное состояние орфографических норм, создает проблемы при лингвистической аннотации текстов и поиске в корпусе. Помимо орфографических вариантов корпус текстов первой половины XX века отличается повышенной степенью вариативности на других уровнях – морфологии, словообразования, синтаксиса. Решить эту проблему предлагается путем расширения словаря за счет внесения в него вариантов, в том числе орфографических.

В ходе выполнения проекта в 2012 году был проанализирован состав несловарных форм по текстам 1-ой пол. XX в., составлена база вариантов в объеме более 1 тыс. единиц. Произведена ручная лемматизация вариантов, отобраны возможные кандидаты для пополнения словаря корпуса. Часть вариантов внесена в файл конфигурации НКРЯ, что должно повысить качество поиска. В процессе анализа вариативности проведена работа по сбору и исправлению ошибок в текстах и в морфологической разметке.

5. Совершенствование системы поиска.

В рамках решения этой задачи, которая предусматривает включение в пользовательский интерфейс дополнительных метатекстовых признаков, подготовлена база данных, включающая расширенные биографические сведения об авторах. Это позволит использовать в отборе подкорпусов такие признаки, как годы жизни авторов, место рождения, род занятий, принадлежность к эмиграции.

6. Пополнение корпуса современной русской прессы, коррекция баланса, метатекстовой и морфологической аннотации.

В 2012 г. в газетный корпус был пополнен текстами, относящимися к 2009, 2010 и 2011 годам. 1) Был осуществлен отбор текстов по заданным параметрам. 2) Проведено редактирование метаразметки и тестирование корпуса, исправление ошибок. В настоящее время общий объем корпуса составляет более 173 млн словоупотреблений. Количественное распределение текстов корпуса по годам выглядит следующим образом.

Период	2000-	2004	2005	2006	2007	2008	2009	2010	2011

	2003								
Объем, млн словоупотр.	29,9	16,4	18,3	18,4	27,8	18,4	19,1	15,5	9,3

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)

Проведен анализ текстового состава корпуса по социологическим и жанрово-стилистическим параметрам, выявлены участки, требующие коррекции баланса. Корпус пополнен текстами в объеме более 4 млн словоупотреблений, которые относятся к разным функциональным сферам, многие из них представляют собой редкие издания. Проведен анализ несловарных форм, составлена база вариантов по текстам 1-ой пол. XX в. в объеме более 1 тыс. единиц. Организована базы данных, содержащая сведения об авторах текстов 1-ой пол. XX в. Проведена коррекция баланса, метатекстовой и морфологической аннотации корпуса современной русской прессы, корпус пополнен текстами за 2011 год, его объем составляет 173 млн словоупотреблений.

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

Проведен анализ текстового состава корпуса по социологическим и жанрово-стилистическим параметрам, выявлены участки, требующие коррекции баланса. Корпус пополнен текстами 1-ой пол. XX в. в объеме более 4 млн словоупотреблений. Проведен анализ несловарных форм, составлена ба-

зы вариантов по текстам 1-ой пол. XX в. в объеме более 1 тыс. единиц. Организоване база данных, содержащая сведения об авторах текстов 1-ой пол. XX в. Проведена коррекция баланса, метатекстовой и морфологической аннотации корпуса современной русской прессы, корпус пополнен текстами за 2011 год, его объем составляет 173 млн словоупотреблений.

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 системное пополнение корпуса 1-ой пол. XX в. с учетом проведенного анализа текстового состава корпуса. Ожидаемый объем нового материала - не менее 4 млн словоупотреблений. Будет продолжен анализ несловарных форм, пополнение базы вариантов по текстам 1-ой пол. XX в. в объеме до 1 тыс. единиц. Будет пополнен словник морфологического словаря и коррекция опечаток и ошибок аннотации. В базу авторов будут включены новые данные на основе текстов корпуса. Планируется системное пополнение корпуса современной русской прессы в объеме до 20 млн словоупотреблений.

Подпись руководителя проекта А.Я. Шайкевич