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 основного корпуса современных текстов НКРЯ</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Савчук Светлана Олеговна, к.ф.н</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>savsvetlana@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="7">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>Летучий А.Б., кфн, ИРЯ РАН</td></tr>
<tr><td>Ловля Е.Н., ИРЯ РАН</td></tr>
<tr><td>Сердобольская Н.В., кфн РГГУ</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 Системное пополнение основного корпуса современных текстов НКРЯ

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)
Савчук Светлана Олеговна, к.ф.н., с.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, Сердобольская Н.В., к.ф.н., доц. МППГУ, ИРЯ РАН, Сичинава Д.В., кфн, ИРЯ РАН, Летучий А.Б., кфн, ИРЯ РАН, Ловля Е.Н., ИРЯ РАН, инженер, Морозова Е.Н., ИРЯ РАН, инженер, Алексеевский Л.Д., ФГУ РНЦ «Курчатовский институт», инженер-программист

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

Национальный корпус современного языка по определению не может содержать всех текстов, созданных на этом языке. Но он должен быть представительным, т. е. с максимальной полнотой отражать современное речевое употребление во всем разнообразии функциональных разновидностей и в диахронической перспективе. Настоящий проект направлен на системное пополнение текстового состава корпуса с базовой разметкой. Пополнение уже функционирующего корпуса предполагает предварительный отбор текстов по заданным критериям и в определенном соотношении, опирающемся на анализ существующего баланса текстов. Самую значительную по объему

часть базового корпуса составляют основной корпус современных текстов (условно с середины XX в. до наших дней).

Анализ текстового состава корпуса по состоянию на 2012 год выявил следующую картину. Распределение текстов основного корпуса по сферам коммуникации и по периодам представлено в таблицах 1 и 2.

Таблица 1.

сфера	**1951-2000**	**%**	**НКРЯ в целом**	**%**
всего с/у	67 252 763	100%	209 203 107	100%
художественная	35 738 619	53%	92 060 453	44%
публицистика	23 363 510	35%	76 002 916	36%
учебно-научная	4 873 391	7%	27 375 161	13%
официльно-деловая	773 661	1,2%	3 854 693	1,8%
бытовая	456 998	0,6%	4 443 085	2%
церк-богословская	1 463 126	2,1%	3 222 162	1,5%
произв-техническая	679 339	1%	1 326 442	0,6%
реклама	73 694	0,1%	572 414	0,3%
электр. коммуникация	0	0	2 286 557	1%

Таблица 2

сфера	**2001-2004**	**%**	**2005-1011**	**%**
всего с/у	43 751 508	100%	7 155 699	100%
художественная	8 545 487	20%	2 917 768	40%
публицистика	25 103 601	57%	2 158 391	30%
учебно-научная	7 508 516	17%	417 430	6%
официльно-деловая	1 073 936	2,5%	12 640	0,2%
бытовая	168 464	0,4%	25 717	0,3%
церк-богословская	118 518	0,4%	0	0
произв-техническая	233 658	0,5%	0	0
реклама	430 088	1%	49 249	0,6%
электр. коммуникация	767 392	1,7%	1 557 932	22%

Как видно из таблицы 1, сравнение корпуса 2-й пол. XX в. с данными корпуса в целом показало, что для 2-й пол. XX в. наибольшее отклонение от средних значений наблюдаются у текстов учебно-научной сферы, бытовой и рекламы (эти значения вдвое ниже, чем в среднем по корпусу). Поэтому с целью устранения обнаруженных диспропорций в ходе выполнения проекта в 2012 г. в архив корпуса были включены научные и научно-популярные тексты в объеме 30 Мб, которые будут подготовлены и размещены в корпусе в 2013 г.

Таблица 2 показывает распределение по сферам коммуникации внутри корпуса текстов XXI в. Этот подкорпус является значительным по объему (более 50 млн словоупотреблений) и в целом неплохо сбалансирован. Так, соотношение текстов, относящихся к основным сферам коммуникации, выглядит следующим образом:

- Художественная – 22,4%
- Публицистическая – 53,5%
- Учебно-научная – 15,7%

- Официально-деловая – 2,3%
- Бытовая – 0,4%
- Электронная коммуникация – 4,6%
- Другие – 1,1%

Ряд отличий от средних значений по корпусу в целом объективно отражает, как представляется, те изменения, которые наблюдаются в функционировании текстов в XXI в. Так, изменение соотношения художественной литературы и публицистики в пользу последней отражает тот поворот читательского интереса к литературе нон-фикшн (мемуары, дневники, путевые записки, эссе и пр.), который неоднократно отмечался социологами, аналитиками, литературными критиками. Слабо представлена в корпусе бытовая сфера – на фоне резкого роста доли электронной коммуникации, что также отражает реалии сегодняшнего дня, когда традиционные бытовые жанры – записка, письмо, дневник – почти целиком переместились в интернет и телефонию и успешно там функционируют.

Однако рассмотрение состава текстов XXI в. в хронологическом разрезе показывает, что подавляющая часть текстов относится к первой половине десятилетия, что имеет свое объяснение: формирование основного массива современных текстов происходило в 2004-2006 гг. Поэтому в рамках проекта предстоит наращивать объем текстов новейшего периода, созданных после 2005 г., уделив при этом большее внимание текстам таких функциональных сфер, как научная, производственно-техническая, официально-деловая и церковно-богословская.

В 2012 году работа над проектом велась по нескольким направлениям.

1. **Системное пополнение основного корпуса современных письменных текстов** имело «точечный» характер и состояло в подготовке и включении в корпус недостаточно представленных в нем текстов, относящихся к отдельным жанрам, тематическим областям, временным периодам второй половины XX в. Корпус пополнился большой коллекцией газетно-журнальных текстов (журналы «Огонек» за 1950-1970-е, 1990-е гг., «Сельская новь» за 1988 г., «Общая газета» за 1990-е гг.). Кроме того, были подготовлены и включены в корпус художественные произведения и публицистика авторов, недостаточно в нем представленных (Д. Дар, Л. Кабо, Е. Велтистов, Л. Чуковская, Ю. Трифонов, Ю. Нагибин, Г. Владимов, В. Лихоносов, Ф. Абрамов, А. Твардовский, С. Залыгин, С. Антонов, М. Ганина и др.); научная литература (Н.П. Бехтерева, Ю.А. Жданов, и др.). Общий объем подготовленных текстов составляет около 4 млн словоупотреблений. Выполнение этой задачи будет способствовать коррекции баланса корпуса.

2. **Пополнение основного корпуса текстами новейшей русской прозы (от 2005 г.)**. На основе поведенного мониторинга начат отбор текстов, относящихся к разным сферам функционирования, и формирование архива текстов, объем которого в 2012 г. составил 30 Мб. Осуществлено редактирование электронных версий текстов, снабжение их метатекстовой разметкой. Подготовлены и размещены на сайте НКРЯ тексты, представляющие художествен-

ную прозу разных жанров (Дина Рубина, Виктор Пелевин, Захар Прилепин, Герман Садулаев, Роман Сенчин, Денис Гуцко, Андрей Рубанов, Дмитрий Данилов, Илья Анпилогов, И.В. Бояшов, Мариам Петросян, Сергей Шаргунов, Михаил Шишкин и др.), научно-популярную прозу (журналы «Наука и жизнь», «Детали мира» и др. Объем подготовленного материала – 2 млн словоупотреблений.

3. П**ополнение основного корпуса текстами новейшей русской публицистики (от 2005 г.).** В ходе выполнения этой задачи произведен отбор авторов и текстов по основным тематикам (политика, общество, искусство и культура, право, частная жизнь и т. д.) и сферам коммуникации (пресса, электронные СМИ, блогосфера), объем собранного архива составил 40 Мб. Подготовлены и размещены на сайте тексты объемом 2 млн словоупотреблений: номера журналов «Русский репортер», «Частный корреспондент», электронных СМИ (publiciti.ru и др).

4. Системное пополнение корпуса текстов электронной коммуникации. Расширение корпуса современных текстов за счет текстов электронной коммуникации необходимо потому, что именно эта сфера, наряду с устной речью, непосредственно и чутко реагируют на перемены в общественной жизни. Отбор текстов для корпуса опирается на предварительный мониторинг сайтов по таким параметрам, как тематические области, жанровый состав текстов, социологические характеристики участников общения (примерный возраст, профессиональный и образовательный уровень). Это позволяет поддерживать баланс текстов и охватить широкое тематическое пространство. В ходе выполнения проекта в 2012 г. в корпус были включены форумы и блоги широкого тематического спектра (образование, досуг, зрелища и развлечения, медицина и здоровье, частная жизнь, транспорт, путешествия и др.). Тексты электронной коммуникации, отражающие непосредственное неформальное общение, отличаются свободным отношением к орфографической норме и высокой степенью вариативности, вследствие чего их подготовка требует проведения ряда дополнительных процедур (нормализации орфографии и социологической аннотации). По степени сложности она приближается к подготовке транскриптов устных текстов. Общий объем подготовленного материала составил 700 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.)

Савчук С.О. Варианты родовой принадлежности в группе существительных pluralia tantum в русском языке // Компьютерная лингвистика и интеллектуальные технологии. Вып. 11 (18). М., 2012. Т.1. С. 548-558

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 мониторинг текущего состояния корпуса современных текстов с базовой разметкой, определена стратегия его пополнения. Проведено системное пополнение основного корпуса: в него включены тексты 2-й пол. XX в. в объеме 4 млн словоупотреблений, относящиеся к художественной, публицистической и научной сферам, отдельным жанрам, тематическим областям, временным периодам второй половины XX в. Проведено пополнение основного корпуса текстами новейшего периода. В корпус включены произведения художественной и научной прозы в объеме 2 млн словоупотреблений. Подготовлены тексты современной публицистики (журналы и электронные СМИ) в объеме 2 млн словоупотреблений. Произведено системное пополнение корпуса текстов электронной коммуникации в объеме 0,7 млн словоупотреблений.

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

В 2012 г. проведено системное пополнение основного корпуса: в него включены тексты 2-й пол. XX в. в объеме 4 млн словоупотреблений; художественная и научная проза новейшего периода (после 2005 г.) в объеме 2 млн словоупотреблений и современная публицистика в объеме 2 млн словоупотреблений. Корпус текстов электронной коммуникации пополнен в объеме 0,7 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 году планируется системное пополнение основного корпуса: тексты 2-й пол. XX в. в объеме 2 млн словоупотреблений (прежде всего научные

тексты, а также отдельные авторы, недостаточно представленные в корпусе). Корпус текстов новейшего периода будет пополнен в объеме 5 млн словоупотреблений текстами художественной прозы, в том числе жанровой (детская литература, фантастика, историческая проза), научными и научно-популярными, публицистическими текстами. Будут подготовлены тексты, относящиеся к официально-деловой и церковно-богословской сферам. Планируется системное пополнение корпуса текстов электронной коммуникации в объеме до 0,7 млн словоупотреблений.

Подпись руководителя проекта С.О. Савчук

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
	Системное пополнение основного корпуса современных текстов НКРЯ	ИРЯ РАН	С.О. Савчук, кфн		**2013 г.** — Системное пополнение основного корпуса: тексты 2-й пол. XX в. в объеме 2 млн словоупотреблений. Пополнение основного корпуса текстами новейшего периода (художественная и научная проза, публицистика, официально-деловые и церковно-богословские тексты) в объеме 5 млн словоупотреблений. Системное пополнение корпуса текстов электронной коммуникации в объеме до 0,7 млн словоупотреблений.