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
1. Создание и развитие корпусных ресурсов по современному русскому языку</td></tr>
<tr><td colspan="2">Название проекта
Создание словарного модуля Национального корпуса русского языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Плунгян Владимир Александрович, чл.-корр. РАН</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>plungian@gmail.com</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>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
119019 Москва, ул. Волхонка 18/2
тел.: (495) 695 26 60, факс: (495) 695 26 03
ruslang@ruslang.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="3">Исполнители (ФИО, уч. ст.)</td><td>Кустова Г.И., д.ф.н., МПГУ</td></tr>
<tr><td>Ляшевская О.Н., к.ф.н., ИРЯ РАН</td></tr>
<tr><td>Мазурова Ю.В., к.ф.н., ИЯ РАН</td></tr>
<tr><td></td><td>Савчук С.О., к.ф.н., ИРЯ РАН</td></tr>
<tr><td></td><td>Сичинава Д.В., к.ф.н., ИРЯ РАН</td></tr>
<tr><td></td><td>Бирюк О.Л., ИЯ РАН</td></tr>
<tr><td></td><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 1. Создание и развитие корпусных ресурсов по современному русскому языку

2. Название проекта Создание словарного модуля Национального корпуса русского языка

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы) Плунгян Владимир Александрович, чл.-корр. РАН, зав. отделом корпусной лингвистики и поэтики ИРЯ РАН

4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы) Кустова Г.И., д.ф.н., МПГУ, профессор; Ляшевская О.Н., к.ф.н., ИРЯ РАН, докторант; Мазурова Ю.В., к.ф.н., ИЯ РАН, н.с.; Савчук С.О., к.ф.н., ИРЯ РАН, с.н.с.; Сичинава Д.В., к.ф.н., ИРЯ РАН, н.с., Бирюк О.Л., ИЯ РАН, асп., Поляков А.Е., НПБ им. К.Д. Ушинского

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

Выполнение проекта ведется по двум направлениям: 1) создание грамматического словаря на основе текстов XVIII в. и 2) разработка электронного словаря новых слов на основе корпуса современных текстов.

I. **Электронный грамматический словарь на основе текстов XVIII в.**

Создание электронного грамматического словаря должно обеспечить принципиальное улучшение качества морфологического анализа текстов предшествующих периодов, в том числе в дореволюционной орфографии. Разработка словаря предполагает, во-первых, формирование словника на основе существующих лексикографических источников, а во-вторых, пополнение словника единицами, отсутствующими в словарях и извлеченными непосредственно из корпуса текстов. Расширение словаря и настройка морфологического анализатора на определенные тексты позволит в дальнейшем повысить

качество автоматической морфологической разметки не только текстов XVIII в., но и всего диахронического модуля.

В 2012 году были решены следующие задачи.

1) *Создание пилотной версии словника на базе существующих словарей.*

Для решения этой задачи были использованы электронные версии словарей, из которых были выделены и особым образом обработаны словники. Были использованы следующие издания:

– Словарь церковнославянского и русского языка. – СПб., 1847 (~114 тыс. слов);

– Словарь Академии Российской. Т. 1–6. – СПб., 1789–1794 (~49 тыс. слов и выражений);

– Словарь русского языка 18 века. Т. 1–14. – СПб., 1984–2004.

Кроме того, в качестве сравнительной базы также использовался словник «Грамматического словаря современного русского языка» А. А. Зализняка (~110 тыс. слов).

В результате сравнения и анализа разных словарей было выявлено, что наиболее перспективно делать сводный словник из словаря Зализняка и Словаря церковнославянского и русского языка (СЦРЯ). Словарь Зализняка содержит наиболее полную информацию о грамматике и словоизменении, а также является источником для построения гипотез для неизвестных слов. СЦРЯ достаточно хорошо отражает лексику конца XVIII–XIX веков и также содержит достаточно подробную грамматическую информацию, хотя и в другом формате.

Несмотря на то, что пересечение словников обоих словарей составляет не более 50% (~46 тыс. слов), сюда входят наиболее частотные слова и морфологические модели. В результате слияния обоих словарей был получен сводный словник на 164 тыс. слов, который далее был обработан вручную с целью выявления сходств и различий между языком XVIII–XIX века и современным.

2) *Анализ словника и приписывание грамматической информации и парадигм*

Эта работа проводилась на основе грамматической информации, содержащейся в словарях. Для слов, совпадающих с современными, приписывалась грамматическая информация из словаря Зализняка (с соответствующими коррективами), а для остальных слов была взята неполная грамматическая информация из СЦРЯ (с соответствующими коррективами).

артиллерии - артиллерия=N,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37

гремит - греметь=V,ipf,intr=fin,ind,pres,sg,3,act=V5p

алчущу - алкать=V,ipf,intr=ptcp,pres,act,short,sg,f,acc,old,act=V6t | алчущий=A=short,sg,f,acc,old=A4

глаголати - глаголать=V,ipf,tr=inf,old,act=V6n

В настоящее время выполняется унификация обеих частей словника и дополнение отсутствующей грамматической информации во второй части. В данный момент полностью проверено около 30 тыс. лексем.

3) Анализ частотного списка словоформ

В результате анализа реальных текстов для корпуса XVIII века был получен список, содержащий 256 тыс. различных словоформ. Вначале список был обработан с помощью морфологического анализатора Mystem, который показал не очень хорошие результаты – множество устаревших форм оказались неразобранными или для них были построены нелепые гипотезы.

В качестве альтернативы список был обработан морфологическим анализатором А.Е. Полякова, специально адаптированным для устаревших форм, который породил значительно более правдоподобные разборы и гипотезы. В числовом выражении результаты выглядят следующим образом: разобрано 185221 словоформ, что составляет 72.4% всех форм; построены гипотезы для 63904 словоформ, что соответствует 25.0% всего списка; не разобрано 6780 форм, или 2.6% форм.

В настоящее время проводится ручная проверка результатов работы парсера с целью его усовершенствования. Основная проблема при анализе текстов XVIII века – это устаревшие лексемы, отсутствующие в современных словарях. Хотя парсер обычно порождает для них разумные гипотезы, было бы лучше ввести наиболее частотные из них в словарь анализатора.

Некоторую проблему составляют устаревшие формы, особенно церковнославянские, которые регулярно встречаются в некоторых текстах высокого штиля (Прокопович, Бужинский и др.). Видимо, имеет смысл ввести наиболее частотные церковнославянские формы в грамматические таблицы парсера.

В перспективе предполагается объединить оба направления работы – словарную и анализ текстов. Нужно, чтобы морфологический анализатор использовал сводный словник, включающий СЦРЯ (см. выше), а также пополнить словарь анализатора частотными словами из текстов.

С целью отбора кандидатов на пополнение словаря была проведена оценка частотного списка единиц, получивших гипотетические грамматические разборы. Основным критерием отбора является частотность словоформы. Так, вряд ли необходимо включать в электронный словарь корпуса низкочастотные имена собственные (такие как *Азбад, Айшедуд, д'Аркур, Ардильер, Блюмфельберг* и под.), однако высокочастотные в текстах XVIII–XIX в. слова должны в нем присутствовать (*вышеписанный, вышепомянутый, государствование, доношение, деташемент, апробовать, аще, егда* и под.). Другим немаловажным критерием является формальный облик слова: если словоформа не дает возможности более или менее точно предсказать его лемму и грамматические характеристики, слово с этой информацией следует поместить в словарь. В особенности это касается архаизмов (*биет, благословляяй, болезнех, бысть, варвари, вещми* и др.). Так для формы *агнчий* предлагается 12 разборов, из которых только 2 правильных, для *ангельстии* - 16 вместо 1, для *алкаличных* - 11 вместо 3, для *англяне* - 24 разбора вместо 1 правильного, и т.д. Наличие вариантов также является основанием для помещения слова в словарь, а если какой-либо вариант уже присутствует в словаре, то новый приписывается с помощью соответствующих ссылок.

4) Выделение орфографических, фонетических и морфологических вариантов, характерных для текстов XVIII века

Работа по выделению вариантов в текстах XVIII века ведется на основе экспериментального корпуса текстов XVIII - начала XIX в. общим объемом около 4 млн словоупотреблений. Проведена обработка текстового массива, получены частотные списки словоформ, содержащие сведения об их грамматических характеристиках. Построен частотный словарь словоформ, не получающих предсказанного разбора, эти словоформы снабжены гипотетическими разборами, правильными или ошибочными. Организована база данных несловарных словоформ, имеющая следующие поля: словоформа, сгенерированная лемма, отсылочная (нормализованная) лемма, грамматические признаки леммы, грамматические признаки словоформы, тип варианта, сведения об авторе и дате создания текста, в котором зафиксирована форма.

В 2012 году проведен анализ 4000 единиц базы данных: для каждой словоформы приписана или выбрана из предложенного списка правильная лемма, приписаны отсылочная (нормализованная) словоформа и нормализованная лемма, грамматическая информация проверена по корпусу и определены правильные грамматические характеристики. В процессе анализа базы данных несловарных словоформ выявляются варианты, характерные для текстов XVIII в., проводится их разметка. Выделены орфографические, фонетические и морфологические варианты, не учтённые в грамматическом словаре НКРЯ. Среди них самую заметную часть составляют орфографические варианты – около 76% (*брегодир, верьху, ветьви, весма, болнова, блиский, выш-ший*), которые бывает трудно характеризовать отдельно от фонетических (*император*, *апотеоз, апеллации*, *арриергард, арьерград, аувстрийский, билиары, билиарды, бонбы, Выбурх* и под.). Морфологические варианты составляют около 2% (*бедствиев* вм. *бедствий, венгерцов, вероломцов, владельцов* вм. *венгерцев, вероломцев, владельцев; белилы, блюды* вм. *белила*, *блюда*, *балюстрадом* вм. *балюстрадой* и др.); словообразовательные - 18% (*венецианский, веницейский, великость, кислость, бедство, бездельство, благодарствие, давный*, *дальный*, *присвоять, остановлять, благоустроивать, присвоивать* и под.).

II. Электронный словарь новых слов на основе корпуса современных текстов

1. Анализ словоформ

В ходе работ по Проекту проводился дальнейший анализ словоформ, присутствующих в современных текстах, но не получающих адекватного разбора. Всего обработана база на 20 тысяч вариантов анализа. При этом использовались следующие источники сведений:

а) частотные списки словоформ, получающих гипотетический разбор. Анализировались слова частотой 10 и ниже.

б) списки словоформ, получающих максимум гипотетических разборов. Морфологически анализатор Mystem, использующийся в Корпусе, выдаёт

особо большое количество вариантов (иногда до 50-60) для слов на *и* и с рядом других исходом (например, неизменяемое существительное в 12 формах, императив, многочисленные падежные формы существительных III склонения, формы прошедшего времен глагола на *-ли* и т. п.)
в) сообщения пользователей об ошибках автоматического морфологического разбора в Корпусе.
г) список расхождений между морфологическими разборами корпуса со снятой омонимией и автоматического анализатора.
Часть выделяемых словоформ характерна для церковнославянского языка, диалектизмов и языка XVIII века, поэтому отчасти пересекается с формами, выделенными выше; часть – неологизмы, в том числе частотные в текстах электронной коммуникации современные жаргонизмы *(аська, горбачевский, офф, хрень),* имена собственные, слова, встретившиеся в добавленных в НКРЯ за 2010-2011 гг. текстах. Значительная часть добавляемых слов – неизменяемые наречия, предикативы, частицы, предлоги, междометия *(*слова типа *ино, повдоль, намелко, -тка*, *хныкать* и т. п.*),* для которых сложно предсказать часть речи на основании правой части словоформы.

2. Пополнение словаря и внедрение его в разметку НКРЯ
В результате анализа списку «проблемных» словоформ приписывался набор отсутствующих в выходе морфологического анализатора разборов, в том числе с пометой anom (аномальная форма). В дальнейшем этот список был внедрён в файл конфигурации морфологического анализатора Mystem и используется при автоматической разметке Корпуса начиная с релиза 2012 г. Данный словарь применяется ко всем русским текстам Корпуса (более 400 миллионов словоупотреблений), кроме текстов со снятой омонимией (разметка которых корректируется отдельно вручную).
Всего введены индивидуальные правила для 1000 словоформ, таким образом, файл конфигурации (словарный дополнительный модуль) анализатора вырос почти вдвое.
Впервые предусмотрена возможность отдельного анализа форм, пишущихся с прописной и строчной буквы. Это позволяет лучше отграничить разбор имён собственных и аббревиатур от омографичных с точностью до регистра имён нарицательных (например, *ПО* в значении «программное обеспечение» и предлог *по*, *путина* и *Путин* и т. п.), и, соответственно, сократить омонимию.
Пример формата индивидуальных разборов в файле конфигурации:

```
ехалось <ana lex="ехаться" gr="V,ipf,intr=praet,n,sg,indic"/>
едется <ana lex="ехаться" gr="V,ipf,intr=praes,3p,sg,indic"/>
евро <ana lex="евро" gr="S,m,inan,0=(sg,pl),(nom|gen|dat|acc|ins|loc)"/><ana lex="евро" gr="S,n,inan,0=(sg,pl),(nom|gen|dat|acc|ins|loc)"/>
+драла <ana lex="драла" gr="PRAEDIC"/>
+драло <ana lex="драло" gr="PRAEDIC"/>
расшиши <ana lex="расшиши" gr="S,pl=(nom|acc)"/>
штандер <ana lex="штандер" gr="S,m,inan=(nom|acc)"/>
квиты <ana lex="квиты" gr="PRAEDIC"/>
```

квит <ana lex="квит" gr="PRAEDIC"/>
влом <ana lex="влом" gr="PRAEDIC"/>
+прикольно <ana lex="прикольно" gr="PRAEDIC"/>
+пятистенок <ana lex="пятистенок" gr="S,m,inan=sg,(nom|acc)"/>
+пятистенке <ana lex="пятистенок" gr="S,m,inan=sg,loc"/>
+пятистенка <ana lex="пятистенок" gr="S,m,inan=sg,gen"/>
словей <ana lex="слово" gr="S,n,inan=pl,gen=anom"/>

3. Введение новых правил для автоматической разметки НКРЯ

Кроме того, при автоматической разметке корпуса были внедрены очередные правила разбора целых классов словоформ (в основном это грамматические и орфографические варианты):
характерные для XVIII в. (а затем для имитации «неграмотной» речи) словоформы на *-тца* получает разбор такой же, как с *-ться* и *-тся* (с добавлением пометы distort)
церковнославянские словоформы на *-ши* -- такой же, как на *-шь* (с добавлением anom)
Для форм сравнительной степени: *-ае = -ее*, *-яе = ее*. (например, *сильняе, горчае*)
Приравнивание орфографических вариантов: *цы = ци*.
Разбор орфографически искажённых слов с растягиванием букв и дефисами (запись с помощью регулярных выражений): *а+ = а, а-а(-а)* = а, па-па = папа* (правила могут применяться одновременно: *ааа-ппа-а-а-а = папа*), с добавлением специальной пометы distort
Разбор продуктивных наречий: *по-*(дефис)*...му* (например, *по-школьному*) и *по-*(дефис)*...ки* (например, *по-крымски*) объединяются в единый дефисный разбор, а не разрываются по дефису.
Внедрение этих правил помогает правильно размечать открытые продуктивные классы словоформ, не предусматривая каждое их возможное вхождение в словаре. Ещё более расширяется возможность поиска «аномальных» (не предусмотренных нормами, например, Грамматического словаря) форм по текстам с неснятой омонимией (ранее помета об аномальности формы присутствовала лишь в текстах, где омонимия снималась вручную).
Работа в данном направлении существенно оптимизирует поиск по НКРЯ, снимает уровень «шума» в выдаче, стимулирует исследование лексики и грамматики, в частности, пополнение словарей и грамматических описаний. Далеко не все выявленные формы, особенно орфографические, учтены существующими словарями.

6. Общее число опубликованных в 2012 г. по проекту работ
6.1. количество монографий -
6.2. количество сборников статей -
6.3. количество статей - 4

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.)

Поляков А.Е. Лемматизатор для дореформенной русской орфографии // Информационные технологии и письменное наследие: Материалы IV международной научной конференции (Петрозаводск, 3-8 сентября 2012 г.). Петрозаводск; Ижевск, 2012. С. 208-211

Савчук С.О. Электронный словарь вариантов на основе текстов 18 века // Информационные технологии и письменное наследие: Материалы IV международной научной конференции (Петрозаводск, 3-8 сентября 2012 г.). Петрозаводск; Ижевск, 2012. С. 241-244

Д. В. Сичинава. Национальный корпус русского языка как инструмент лексической типологии // Труды конференции LEXT-III, Гранада, Испания (в печати, сдано 1 ноября 2012 г.)

Sichinava, D. Korpusy równoległe w Narodowym Korpusie Języka Rosyjskiego // Prace Filologiczne.. Seria językoznawcza Tom LXIII, 2012.

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)

Создана основа для электронного грамматического словаря – сформирована пилотная версия словника на базе трех словарей XVIII века, проведен анализ словника и приписывание грамматической информации и парадигм (обработано 30000 лексем). Проанализирован частотный список словоформ, извлеченных из корпуса текстов XVIII-XIX в., наиболее частотных лексемы отобраны для пополнения совокупного словника. На основе анализа базы данных вариантов выделены орфографические, фонетические и морфологические варианты, характерные для текстов XVIII в., которые будут учтены в словаре.

Проведён анализ базы данных новых словоформ по Корпусу (20 тыс. вариантов анализа – архаизмы, неологизмы, имена собственные из редких текстов), пополнен словарный файл конфигурации морфологического анализатора Mystem (1000 индивидуальных разборов), введены новые регулярные правила разбора, они внедрены в морфологическую разметку всех русских текстов НКРЯ с неснятой омонимией. Работа в данном направлении существенно оп-

тимизирует поиск по НКРЯ, снимает уровень «шума» в выдаче, стимулирует исследование лексики и грамматики, в частности, пополнение словарей и грамматических описаний. Далеко не все выявленные формы, особенно орфографические, учтены существующими словарями.

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

Сформирована пилотная версия словника на базе трех словарей XVIII века, проведен анализ словника и приписывание грамматической информации и парадигм (обработано 30000 лексем). Для пополнения совокупного словника отобраны наиболее частотных лексемы, извлеченные из корпуса текстов XVIII-XIX в. Выделены варианты, характерные для текстов XVIII в., которые также будут учтены в словаре.

Проведён анализ базы данных новых словоформ по Корпусу (архаизмы, неологизмы, имена собственные из редких текстов – всего 20 тыс. вариантов анализа), пополнен словарный файл конфигурации морфологического анализатора Mystem (1000 индивидуальных разборов), введены новые регулярные правила разбора, которые внедрены в морфологическую разметку всех русских текстов НКРЯ с неснятой омонимией. Работа в данном направлении существенно оптимизирует поиск по НКРЯ, стимулирует исследование лексики и грамматики, в частности, пополнение словарей и грамматических описаний.

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 обеспечить расширение словника расширение словника грамматического словаря до ~50 тыс. лексем. Будет продолжена обработка словника, коррекция грамматических разборов и приписывание парадигм. Другим источником пополнения словаря будет служить корпусные базы данных - база несловарных словоформ XVIII в. и база вариантов, заполнение которых будет продолжено в 2013 году. Планируется создание пилотной версии морфологического анализатора и тестирование ее на разных корпусах. В процессе конструкторских работ будет производиться сбор данных для коррекции текстов и морфологической разметки диахронического корпуса.

На основе корпусов современных текстов планируется обработка базы данных новых слов на 20 тыс. лексем.

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
	Создание словарного модуля Национального корпуса русского языка	ИРЯ РАН	В. А. Плунгян		**2013 г. –** расширение словника (до ~50 тыс. лексем); создание базы несловарных словоформ XVIII в.; пополнение словника грамматического словаря вариантами XVIII в; создание пилотной версии морфологического анализатора; сбор данных для коррекции текстов и морфологической разметки диахронического корпуса. Обработка базы данных новых слов на 20 тыс. лексем.