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
Создание и развитие параллельных русско-
иноязычных корпусов в Национальном корпусе русского языка</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Добровольский Дмитрий Олегович, дфн</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>119019 Москва, ул. Волхонка 18/2
dm-dbrv@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>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Глебова Татьяна Николаевна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования</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., ИРЯ РАН</td></tr>
<tr><td>Сичинава Д. В., ИРЯ РАН</td></tr>
<tr><td></td></tr>
<tr><td>Дата сдачи отчета 20 ноября 2012 г.</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
**Создание и развитие параллельных русско-
иноязычных корпусов в Национальном корпусе русского языка**
3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)
Добровольский Дмитрий Олегович, доктор филологических наук, ведущий научный сотрудник, зав. Отделом ИРЯ РАН
4. Основные участники проекта (ФИО *полностью*, ученая степень, должность, место работы)
Савчук Светлана Олеговна, кандидат филологических наук, старший научный сотрудник ИРЯ РАН
Сичинава Дмитрий Владимирович, кандидат филологических наук, старший научный сотрудник ИРЯ РАН
5. Содержание фактически проделанной работы, полученные результаты, их новизна, научная и практическая значимость, актуальность, соответствие мировому научному уровню (до 5 стр.)

Разработка и информационное наполнение параллельных корпусов на базе технологий Национального корпуса русского языка

1. Описание процедур и методов сбора и обработки материалов для наполнения банка данных параллельного корпуса

Разработка параллельного корпуса русско-английских и русско-немецких текстов на базе технологий Национального корпуса русского языка была успешно продолжена в отчетный период. Русско-английская часть корпуса бидирекциональна по своей природе. Иными словами, этот компонент параллельного корпуса включает в себя как тексты русского оригинала с переводом на английский язык, так и английские оригинальные тексты с их переводом на русский. Представляется, что создаваемый лингвистический ресурс найдет применение в практике создания словарей, учебных материалов, в подготовке переводчиков, в преподавании иностранных языков и русского языка как иностранного, при состав-

лении и анализе учебников и учебных пособий с точки зрения их соответствия реальному речевому употреблению изучаемого языка.

Выбор включенных в него текстов был мотивирован в первую очередь их высоким литературным качеством, а также наличием квалифицированно выполненных переводов на русский и английский языки соответственно.

В качестве текстового материала для включения в параллельный корпус были отобраны следующие произведения.

Англо-русский компонент параллельного корпуса

- Оскар Уайлд. Портрет Дориана Грея
- Чарльз Диккенс. Записки Пиквикского клуба
- Лемони Сникет. Кровожадный карнавал
- Роберт Луис Стивенсон. Похищенный, или Приключения Дэвида Бэлфура
- Роберт Луис Стивенсон. Катриона
- Джон Голсуорси. Сага о Форсайтах. Сдается в наем
- Джон Голсуорси. Интерлюдия. Последнее лето Форсайта (Сага о Форсайтах-2)
  - Джон Голсуорси. Сага о Форсайтах — 3.
  - -12 рассказов Артура Конан Дойля
  - 12 повестей Л. Фрэнка Баума о стране Оз
  - 10 рассказов О. Генри
  - К. С. Льюис. Конь и его мальчик
  - Любовный роман (Миранда Ли)

В работе

- Джером Д. Сэлинджер. Над пропастью во ржи
- 4 рассказа Артура Конан Дойля

Русско-английский компонент параллельного корпуса

- М.А. Булгаков. Мастер и Маргарита

В работе

- Л.Н. Толстой. Война и мир
- 11 рассказов Л.Н. Толстого
- Максим Горький. Мать
- 4 рассказа Максима Горького

Немецко-русский компонент параллельного корпуса

- Готфрид Август Бюргер. Удивительные путешествия барона Мюнхгаузена
- Фридрих Дюрренматт. Правосудие
- Фридрих Дюрренматт. Подозрение.
- Густав Майринк. Голем
- Томас Манн. Будденброки

В работе

- Теодор Фонтане - Эффи Брист

На следующий год планируется создание русско-немецкого компонента параллельного корпуса. В работе в настоящий момент находятся следующие тексты:

- Н.В. Гоголь. Шинель (в переводе Рудольфа Касснера)
- рассказы А.П. Чехова
- рассказы А.И. Куприна

В целом, объем запланированного на этот год пополнения соответствующих модулей параллельного корпуса выполнен.

Планировалось довести объем английского компонента до 15 млн., а объем немецкого – до 5 млн. словоформ. В настоящее время объем доступного онлайн английского компонента корпуса составляет 17,1 млн. словоформ (на 2,1 млн. больше предполагаемого). Объем находящегося в свободном доступе немецкого компонента равен 2 млн. словоформ, т.е. уступает запланированным показателям. Однако если учесть тексты, находящиеся в работе, запланированные объема окажутся практически выполненными. Некоторое замедление работы с немецким модулем корпуса (по сравнению с ожидаемыми темпами) объясняется несколькими причинами. С одной стороны, немецкие тексты, включая переводы русской литературы на немецкий язык, доступны в электронной форме в гораздо меньшем объеме, чем английские. С другой стороны, более сложная немецкая графика (большое количество диакритических и нестандартных для латинского алфавита знаков, ср. ä, ö, ü, ß) приводит к появлению большого количества сбоев и ошибок при технической обработке текстов, которые удается исправить только вручную, т.е. немецкие тексты должны внимательно вычитываться при редактировании. Эта работа требует больших затрат времени и высокой квалификации редакторов.

Белорусско-русский и русско-белорусский компоненты параллельного корпуса

Объём белорусского корпуса существенно вырос и включает в себя не только культурно значимые тексты на белорусском языке (а также русскоязычных белорусских писателей –

С. Алексиевич, А. Курейчика – в переводе на белорусский), но и тексты законодательных актов, религиозные тексты, газетные статьи.

Конкретнее, в параллельный белорусско-русский корпус входит публицистика (газетные статьи В. Орлова, Ю. Андреевой, Н. Николаевой), юридические документы, художественные произведения XX-XXI в. (Я. Колас, «В полесской глуши», Я. Мавр, «Сын воды», «Полесские робинзоны», «В стране райской птицы», В. Короткевич, «Колосья под серпом твоим», «Седая легенда», «Дикая охота короля Стаха», «Черный замок Ольшанский», «Книгоноши», «Лазурь и золото дня», «Оружие», «В шалаше», «Полещук», «Письма не опаздывают никогда», Я. Брыль, «Нижние Байдуны», В. Быков, «Сотников», «Знак беды», «На болотной стежке», «Карьер», «Волчья яма», «Эстафета», «Одна ночь», «Круглянский мост», «Свояки», «Третья ракета», В. Шитик, «Скачок в ничто», М. Стрельцов, «Голубой вечер», «Двое в лесу», «Четвертый год войны», «Сено на асфальте», «Смаление вепря», А. Кулаковский, «Двенадцатый жёсткий», И. Шемякин, «Торговка и поэт», И. Мележ, «Люди на болоте», В. Чаропка, «Отречение от тьмы», «Победа тени», «На круги своя»).

В параллельный русско-белорусский корпус входят учебники (география, экономика), доклады о состоянии науки в Беларуси, художественные тексты XIX-XX в. (даты создания оригиналов; белоруские переводы датируются второй половиной XX -- началом XXI в.): Булгаков, "Мастер и Маргарита", Куприн, "Гранатовый браслет", Бунин, "Господин из Сан-Франциско", Айтматов, "Буранный полустанок", Алексиевич, "Чернобыльская молитва", Бабель, "Переход через Збруч", Булгаков, "Полотенце с петухом", Булычев, "Похищение Чародея", Фраерман, "Дикая собака динго", Гайдар, "РВС", Куприн, "Олеся", "Поединок", Набоков, "Сказка", "Музыка", "Облако, озеро, башня", "Терра инкогнита", Пришвин, "Журавль", Сейфуллина, "Павлушкина карьера", Соболев, "Соловей", А. Н. Толстой, "Буратино", Троепольский, "Белый Бим Черное ухо", Зощенко, "Баня", Гумилев, "О стихотворных переводах", "Баллады Роберта Саути", Кузмин, "Раздумья и недоуменья Петра Отшельника", Хармс, рассказы, Строцев, "Журавль", Бажов, "Аметистовое дело", Чехов, "Хамелеон", Паустовский, "Шиповник", Гоголь, "Вечера на хуторе близ Диканьки".

Часть русско-белорусских параллельных текстов входит в состав многоязычного подкорпуса параллельного корпуса НКРЯ -- это тексты, переведённые на русский и белорусский с третьих языков ("Маленький принц" А. де Сент-Экзюпери, "Винни-Пух" А. А. Милна, "Алиса в стране чудес" Л. Кэррола).

Украинско-русский и русско-украинский компоненты параллельного корпуса

Объём украинско-русского и русско-украинского корпусов достиг 10 миллионов словоупотреблений и превосходит крупнейший доступный в Интернете корпус украинского языка Лаборатории компьютерной лингвистики Института филологии им. Шевченко. Сюда входят такие тексты, не представленные в других корпусах, как переписка писателей (например, письма Леси Украинки или Василя Стуса), бланки официальных документов, а также, как и в белорусском – художественная литература, публицистика и законодательные тексты.

Польско-русский и русско-польский компоненты параллельного корпуса

Польско-русский и русско-польский параллельные корпуса растут очень динамически, запланированный объём даже несколько перевыполнен). Здесь хорошо представлена польская художественная проза разных периодов, от поколения Сенкевича и Пруса до современных романов. В корпус входят также газетные статьи, законодательство (Уголовный кодекс Республики Польша). Русско-польский параллельный корпус пока не вполне репрезентативен, но в него входит, в частности, такой нестандартный для параллельных корпусов материал, как проза русского Серебряного века (Брюсов и Сологуб) в современных польских переводах.

Новые корпуса (испанский, итальянский, латышский, армянский)

Собраны и подготовлены к разметке новые корпуса НКРЯ – двух романских языков и двух индоевропейских языков ближнего зарубежья (которые оба относятся к «малым» группам индоевропейской семьи). В испанский и итальянский корпус входят как переводы с русского (например, «Анна Каренина» Толстого, «Град обреченный» Стругацких), так и переводы на русский (например, «Улей» Камило Хосе Селы, «Имя розы» Умберто Эко), в латышский и армянский – в большинстве переводы на русский («Белая гвардия», «Тихий Дон» и другие сочинения). Все эти тексты относятся к художественной литературе, задача жанровой репрезентативности пока не ставилась. Для поиска эти тексты не открыты из-за недостаточной разработанности морфологической разметки (в частности, электронных грамматических словарей), которая будет в дальнейшем отлаживаться (а для латышского языка – вероятно, создаваться).

***

В построении данных компонентов корпуса учитывался мировой опыт создания корпусов параллельных текстов. Для этой цели были проанализированы материалы международных конференций и статей в журналах по корпусной лингвистике, посвященных созданию корпусов параллельных текстов: в частности, номера журналов «Corpus linguistics and linguistic theory» и «International journal of corpus linguistics» за последние годы, материалы параллельных корпусов института «ААС – Корпус Австрийской академии» при Австрийской академии наук (Вена) и программы Parasol университета Регенсбурга (Германия) и недавно вышедшая в серии Cambridge Textbooks in Linguistics монография: Tony McEnery and Andrew Hardie «Corpus Linguistics».

Созданная нами концепция корпуса параллельных текстов базируется на принципах, разработанных в рамках Национального корпуса русского языка.

На первом этапе обработки текстов, подлежащих включению в корпус, производилось тщательное редактирование электронных версий этих текстов. Понятно, что при оцифровке текстов могут возникнуть ошибки и неточности самого разнообразного характера: от ошибок в распознавании отдельных букв при сканировании до потери целых кусков текста. Если эти ошибки не будут устранены на первом этапе, это может привести к серьезным трудностям на этапе выравнивания. Так, пропущенный фрагмент текста оригинала или перевода явится причиной сбоя выравнивания всего последующего произведения, поскольку выравнивание осуществляется полуавтоматически на основе гипотезы о наличии принципиального параллелизма между предложениями оригинала и перевода.

На втором этапе обработки осуществляется фрагментирование текстов, то есть их полуавтоматическая разбивка на отдельные предложения. Технической основой данного этапа является разработанная для Национального корпуса русского языка программа полуавтоматического фрагментирования текстов. Далее, на основе соответствующей инструкции, осуществлялась корректура электронных версий фрагментированных текстов параллельного корпуса англо-русских, русско-английских и немецко-русских текстов.

2. Описание технологии выравнивания текстов параллельного корпуса

Выравнивание текстов на английском и русском, а также на немецком и русском языках программными средствами с последующим редактированием представляет собой наиболее технологически сложный и наукоемкий этап создания параллельного корпуса. Для параллельных корпусов НКРЯ были разработаны специальные программные средства выравнивания и создана система управления корпусом, призванная удовлетворить запросы пользователей. Корпус обеспечен программой **ПарТекс** (идея А.А. Кретова, руководство

И.Е. Ворониной, программирование Д. Спесивцева), имеющей на входе два параллельных текста (оригинал и перевод). Выравнивание осуществляется на уровне предложений. Программа выдает на выходе синтезированный текст, в котором последовательно за каждым предложением оригинала следует соответствующее предложение перевода. В таком синтезированном тексте поиск интересующих пользователя слов может осуществляться штатными средствами обычных текстовых редакторов, например, таких, как «Майкрософт Ворд». В программе ПарТекс для поиска может быть «подстрока в строке» и на выход подается текстовый файл, в который входят все пары предложений оригинала и перевода, содержащие заданную последовательность символов. Возможен поиск как английских или немецких, так и русских слов или словосочетаний.

Одна из наиболее существенных трудностей выравнивания заключается в том, что авторское членение текста на предложения и абзацы не всегда выдерживается в тексте перевода. Кроме того, в разных языках (а иногда и разных изданиях) приняты различные способы графического оформления, что иногда затрудняет определение границ предложения в автоматическом режиме. Ср., например, различные способы оформления переходов от прямой речи персонажей к авторским ремаркам. В таких случаях результаты автоматического выравнивания нуждаются в коррекции, осуществляемой вручную. Программа позволяет обнаружить в параллельных текстах асимметрию такого рода.

Для выравнивания остальных корпусов (кроме английского и немецкого) используется программа «Евклид», являющаяся графическим интерфейсом пользователя (GUI) к доступной в открытом режиме программе выравнивания текстов HunAlign.

Входящий стандарт программы – файл в текстовом формате (ANSI или юникод) с любым числом абзацев, разбиений строки, пробелов в начале строк и т. д. По запросу пользователя загружается два текста – оригинал и перевод, после чего они проходят автоматическую предобработку: удаляются пустые строки, лишние пробелы, каждое предложение переносится на новую строку, при этом учитываются основные сочетания знаков препинания и типы границы предложения.

Затем запускается программа HunAlign, при помощи статистического механизма автоматически выравнивающая два текста полностью, и выход её загружается в графический интерфейс программы «Евклид» для ручного постредактирования.

Программа HunAlign, основываясь на ряде статистических весов (длина предложения, совпадение символов, структура знаков препинания и т. п.) приписывает каждой выровненной ею паре предложений определённый коэффициент вероятности выравнивания. Ес-

ли он ниже нуля, значит, выравнивание данных отрезков текста маловероятно. Кроме того, она автоматически склеивает последовательности предложений на одном языке, соответствующие, с его точки зрения, друг другу. Места склейки в выходном формате при этом указываются.

Пары предложений с отрицательным коэффициентом выравнивания и все склеенные последовательности предложений считаются сомнительными отрезками, нуждающимися в первоочередной проверке вручную. Для графической чёткости коэффициенты в сомнительных отрезках выделяются в программе цветом. При помощи специальной опции в программе «Евклид» скрываются все предложения, кроме сомнительных отрезков. Они просматриваются редактором и в случае ошибки выравнивания (а также случаи сохранения в файле лишней информации – например, записанных в текстовом виде выходных данных и т. п.) редактируются вручную при помощи быстрых механизмов программы. Предусмотрены следующие автоматизированные операции (в виде графичеких кнопок при каждой строке, некоторые из которых открывают отдельное диалоговое окно):

добавление пустой строки
удаление пустой строки
перенос предложения в соседнюю строку
перенос части склеенного предложения – т. н. «обрезка») – в соседнюю строку
добавление вручную «обрезков» к предложению
разрезание предложения вручную на две части

При необходимости ближайший контекст сомнительных отрезков может быть раскрыт (двойным щелчком по скрытой строке) без раскрывания всего остального текста. Затем для просмотра и редактирования можно сделать доступным весь остальной текст (без сомнительных отрезков) и отредактировать его по тому же принципу.

Выровненный текст можно сохранить в формате XML (кодировка Юникод), как в окончательном корпусном, так и в промежуточном формате, доступном для дальнейшего редактирования. При сохранении выровненного текста следует указать язык оригинала и перевода. Они будут сохранены в XML в составе соответствующих тегов. В сохранённом XML пары предложений пронумерованы (им сопоставлены уникальные номера) для облегчения ссылок.

Программа «Евклид» позволяет также добавить к каждому сохранённому тексту метатекстовую информацию (метаразметку), которая записывается в качестве отдельной строки в электронную таблицу (документ, доступный для редактирования программами типа Excel). Пользователь заполняет в специальной форме название текста и имя автора в оригинале и переводе, дату создания текста; язык оригинала и перевода сохраняются автоматически исходя из ранее сохранённой информации.

Выровненные тексты в корпусе сохраняются в файлах в формате XML в кодировке Юникод (UTF-8). Выбор данной кодировки позволяет сохранять, кроме букв русского и белорусского алфавитов, также символы с диакритиками в иноязычных фрагментах, научные символы и т. п.

Выровненные пары предложений объединяются при помощи XML-тега <para>. Тег имеет атрибут ID, в который записывается номер предложения.

Предложения объединяются при помощи XML-тега <se>. Тег имеет обязательный атрибут языка, где указан язык текста: <se lang=rus> для русского или <se lang=bel> для белорусского.

Кроме того, **на текущем этапе выполнения проекта в разметку введён факультативный тег loose, который используется для случаев, когда в переводе часть предложения пропущена, добавлена или изменена**, причем это изменение носит очевидно не случайный характер, а диктуется текстологическими факторами (перевод другой версии текста), художественными или идеологическими установками. Значения атрибута следующие:

Добавление: <se lang=rus loose=add>
Пропуск: <se lang=rus loose=omit>
Изменение: <se lang=rus loose=change>

Структура уточненной метаинформации (метаразметки) устроена так. Метаинфармация сохраняется в отдельном файле (электронной таблице) в формате CSV (значения ячеек разделенны при помощи знака «точка с запятой»). Таблица заполняется при помощи соответствующей формы программы «Евклид». Таблица включает в себя следующие поля метаинформации:

1) название текста в оригинале;
2) год создания текста;
3) имя автора в оригинале;
4) год рождения автора;
5) название текста в переводе;
6) имя автора в переводе;
7) имя переводчика (на языке перевода);
8) язык оригинала;
9) язык перевода.

Применение параллельных корпусов

С помощью параллельных корпусов могут быть не только получены интересные результаты в области теоретического языкознания, так как опора на принципиально сопоставимые аутентичные тексты разных языков позволяет выявить часто неожиданные (как квазиуниверсальные, так и специфические) особенности функционирования языковой системы. Корпус параллельных текстов может быть эффективно использован в практике преподавания языков.

Приведем в качестве примера два вопроса, при решении которых обращение к корпусам параллельных текстов представляется разумным.

1. Как ведут себя определенные структуры входного языка (L1) и их соответствия выходного языка (L2) в аутентичных контекстах? Насколько системные характеристики этих структур способны предсказать их поведение в реальном дискурсе? Какие типы контекстов оказываются релевантными для выбора адекватного эквивалента в языке L2? Иными словами, если исследование корпусов показывает, что стандартные «словарные» L2-эквиваленты данной L1-структуры оказываются неприемлемыми в контекстах определенных типов, необходимо выявить релевантные свойства этих контекстов и соответствующим образом переформулировать условия эквивалентности.

2. В каких случаях переводчики предлагают нестандартные решения? С чем это связано? Мотивированы ли отклонения от оригинала субъективными факторами или объективными межъязыковыми различиями между L1 и L2, накладывающими определенные ограничения на способы перевода исходных структур? Если одному и тому же месту оригинала в разных переводах соответствуют различные эквиваленты, встает вопрос о семантических отношениях между этими эквивалентами. Являются они в языке L2 квазисинонимами или же речь идет о различных интерпретациях L1-структуры? Если соответст-

вующие выражения L2 квазисинонимичны, в чем состоят их семантические, прагматические и сочетаемостные различия? Если же речь идет о различных интерпретациях, чем мотивированы отклонения от соответствующих структур оригинала? При каких условиях они могут быть признаны допустимыми?

Современное состояние сравнительной лексикологии, практики составления двуязычных словарей, а также до известной степени и практики преподавания иностранных языков характеризуется ориентацией на сопоставление более или менее изолированных языковых структур. Отрицательным последствием подобной ориентации является недостаточный учет узуса, то есть тех особенностей синтаксического и сочетаемостного поведения единиц языка, которые нельзя объяснить их системными признаками. Так, в принципе известно, что та или иная структура одного языка не может быть во всех контекстах переведена на другой с помощью своего стандартного эквивалента. В определенных контекстах язык L2 традиционно прибегает к другим способам описания соответствующей ситуации. Известно также, что не существует продуктивных правил, по которым можно было бы вывести подобные отклонения от «стандартной эквивалентности» из неких более общих принципов. Единственный способ описания подобных отклонений – это их тщательная фиксация на аутентичном материале.

Корпус параллельных текстов представляет собой наиболее адекватный инструмент для выполнения этих задач. Та или иная языковая структура, интересующая исследователя, может быть найдена во всех представленных в корпусе контекстах с их переводами на соответствующий язык. Поскольку эти эквиваленты также оказываются встроенными в естественные контексты, на основе полученных с помощью параллельного корпуса материалов могут быть сделаны выводы о зависимости выбора эквивалента от типа контекста. Подобные результаты практически всегда расходятся с теми сведениями, которые мы можем почерпнуть из существующих словарей, являясь тем самым нетривиальными.

Важным параметром, по которому языки могут различаться между собой, является степень употребительности определенных выражений. Так, некоторое выражение *A* языка L1 может стандартным образом переводиться на язык L2 с помощью выражения *B* – вполне корректного с точки зрения норм этого языка. Таким образом, выражения *A* и *B* оказываются эквивалентными в рамках системы соответствующих языков. Тем не менее, их функциональная эквивалентность часто представляется неполной. В частности, это имеет место в случае, когда одно из выражений оказывается в своем языке существенно более употребительным, чем его переводной эквивалент – в своем. Такие случаи хорошо прослеживаются на сопоставлении оригинальных текстов с их переводами на другие языки.

Исследование некоторого языкового явления на основе корпуса параллельных текстов может быть по ряду параметров противопоставлено исследованию этого явления на осно-

ве большого корпуса оригинальных текстов. Отличие оригинальных текстов от переводов заключается и в объеме (миллионы слов оригинальных текстов, производимых ежедневно, против относительно небольшого количества текстов, переводимых с иностранных языков), и в природе авторства (т.е. степени оригинальности и творческой свободы при порождении текста), и в культурном контексте (переводные тексты обычно погружены в культуру исходного языка). Все эти факторы обеспечивают различия между исходными и переводными текстами по целому ряду параметров.

Как уже было отмечено выше, одна из наиболее существенных трудностей выравнивания заключается в том, что авторское членение текста на предложения не всегда выдерживается в тексте перевода. В связи с этим возникает, казалось бы, наивный вопрос: зачем переводчики меняют границы предложений? Не проще ли было бы сохранить авторское членение текста? Пытаясь ответить на этот вопрос, мы обнаружили, что этот вид переводческих трансформаций, несмотря на частоту его использования в переводческой практике, является малоизученным. В процессе изучения данной проблемы были получены результаты, которые, возможно, помогут систематизировать наши знания об этих переводческих приемах, а также позволят получить предварительное количественное описание данного явления.

В разметку корпуса **введены специальные теги**, отмечающие несоответствие переводу оригиналу на уровне добавления, пропуска или значимой замены фрагментов текстов см. выше).

6. Общее число опубликованных в 2012 г. по проекту работ
6.1. количество монографий
6.2. количество сборников статей
6.3. количество статей 15

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.)

Добровольский Д.О. Использование корпусов текстов в двуязычной лексикографии // Среди нехоженых путей. Сб. н. статей к юбилею А.А. Кретова. Воронеж: НАУКА-ЮНИПРЕСС, 2012, 14-25.

Д.О. Добровольский. – D. Dobrovol'skij. German-Russian Phraseography: On a New Dictionary of Modern Idiomatics // German as L2: Phraseodidactic studies. Hamburg: Verlag Dr. Kovac (подписано в печать).

Д.О. Добровольский. – D. Dobrovol'skij. Idiome in der Übersetzung und im zweisprachigen Wörterbuch // PARA FRASESPAL2011. Santiago de Compostela, 2012. В печати.

Д.О. Добровольский, И.Б. Левонтина. О синонимии фокусирующих частиц (на материале немецкого и русского языков) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2012». Выпуск 11 (18): В 2 т. Том 1: Основная программа конференции. – М.: Изд-во РГГУ, 2012. С. 138-149.

Д.О. Добровольский. Русские разговорные обращения: к динамике узуса // Проблемы речевого общения / Отв. ред. Н.Н. Розанова. М.: Азбуковник. В печати.

Д.О. Добровольский. – D. Dobrovol'skij. On the semantic structure of idioms // Paremia 2012. В печати.

Д.О. Добровольский. – D. Dobrovol'skij. The notion of "inner form" and idiom semantics // Proceedings of the Sorbonne-Colloqium on Lexical Semantics. В печати.

Д.О. Добровольский, И.Б. Левонтина. – D. Dobrovol'skij, I. Levontina. Russian NET vs. German NEIN 'NO': a semiotic approach // Russian linguistics (2012), 36. P. 213-219.

Д. О. Добровольский. – D. Dobrovol'skij. Phraseology: Historical Development and Theoretical Aspects // II Congreso Internacional de Fraseología y Paremiología (CIFP). В печати.

Д.О. Добровольский. – D. Dobrovol'skij. „Poznań wart poznania" – Phraseme in den Medien // Studia Germanica Gedanensia 26. Gdansk: WUG, 2012. В печати.

Д. В. Сичинава, Т. А. Архангельский. Русско-белорусский и белорусско-русский параллельные корпуса: совместный проект Национального корпуса русского языка // Труды международной конференции TEL, Казань, КГУ/ издательство «Фэт», 2012.

Д. В. Сичинава. Параллельные корпуса Национального корпуса русского языка как инструмент лексической типологии // Труды симпозиума по лексической типологии LEXT-III, Гранада, Испания (в печати, сдано 1 ноября 2012 г.)

Д. В. Сичинава. Частицы *было* и *бывало*: русские «вторичные модификаторы» в свете типологии и диахронии // Типология славянских балтийских и балканских языков. Спб., Алетейя (в печати, последняя корректура в октябре 2012 г.), 175—194

Д. В. Сичинава. Русская конструкция с *было* и белорусский плюсквамперфект: сопоставительный анализ на материале белорусско-русского и русского-белорусского параллельных корпусов // Компьютерная лингвистика: научное направление и учебная дисциплина. Гомель, ГГУ им. Ф. Скорины, 2012.

Д. В. Сичинава – Dmitri Sitchinava. 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)

Объём параллельных корпусов существенно увеличен, планы по объёму выполнены. Перевыполнены планы по объёму для англо-русского корпуса (17 миллионов словоупотреблений). Параллельные корпуса с участием белорусского и украинского языков стали фактически крупнейшими публично доступными корпусными ресурсами для этих языков, сюда входят многие культурно значимые тексты этих языков. Тексты разнообразных жанров и временных периодов пополнили польско-русский корпус. Использовались разные программные средства (КоПарТ, Евклид). Впервые применена технология разметки неточности перевода (украинский, белорусский, романские корпуса, некоторые польские тексты) Собраны и подготовлены к разметке корпуса романских языков (итальянского и испанского) и индоевропейских языков малых групп – латышского и армянского.

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

Увеличены объёмы англо-русского, немецко-русского, украинско-русского и польско-русского параллельных корпусов, доступных для поиска. Составлены и подготовлены к размещению в Интернете новые параллельные корпуса – итальянский, испанский, армянскй и латышский.

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 ожидается существенное пополнение всех параллельных корпусов: английского (до 20 млн), немецкого (до 10 млн), украинского (до 15 млн), белорусского (до 10 млн), польского (до 4 млн), итальянского (до 3 млн), испанского (до 3 млн), армянского (до 3 млн), латышского (до 3 млн). Также планируется открытие новых параллельных корпусов для поиска (итальянского, испанского, армянского и латышского). Будут также разработаны французский и литовский корпуса. Усовершенствование программных средств (возможно, с подключением словарей) поможет убыстрить и уточнить процесс автоматического выравнивания.

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 и развитие параллельных русско-иноязычных корпусов в Национальном корпусе русского языка	ИРЯ РАН	Д. О. Добровольский + 2 исполнителя		Пополнение английского (до 20 млн), немецкого (до 10 млн), украинского (до 15 млн), белорусского (до 10 млн), польского (до 4 млн), итальянского (до 3 млн), испанского (до 3 млн), армянского (до 3 млн), латышского (до 3 млн) корпусов. Создание французского и литовского корпусов по 2 млн каждый