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 языкам мира»</td></tr>
<tr><td colspan="2">Название проекта «Генеральный корпус современного монгольского языка, версии 2а-2в»</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Крылов Сергей Александрович, доктор филологических наук</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>krylov-58@mail.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования
Институт востоковедения РАН (ИВ РАН)</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Наумкин Виталий Вячеславович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Батяркина Татьяна Владимировна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования ул. Рождественка, д. 12, (495)6211884, факс (495)6211884, inf@ivran.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>Пюрбеев Григорий Церенович, д. ф. н., ИЯ РАН</td></tr>
<tr><td>Петрова Мария Павловна, к. ф. н., Вост. ф-т СПбГУ</td></tr>
<tr><td>Яхонтова Наталья Сергеевна, к. ф. н., Ин-т восточных рукописей 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 **Создание и развитие корпусных ресурсов по языкам мира**

2. Название проекта **Генеральный корпус современного монгольского языка, версии 2а-2в**

3. Руководитель проекта (ФИО *полностью*, ученая степень, должность, место работы)

Крылов Сергей Александрович, доктор филологических наук, ведущий научный сотрудник Отдела языков Азии и Африки ИВ РАН

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

Крылов Сергей Александрович, доктор филологических наук, ведущий научный сотрудник ИВ РАН

Пюрбеев Григорий Церенович, доктор филологических наук, главный научный сотрудник ИЯ РАН

Петрова Мария Павловна, кандидат филологических наук, доцент Восточного факультета СПбГУ

Яхонтова Наталья Сергеевна, кандидат филологических наук, старший научный сотрудник Института восточных рукописей РАН

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

Создана версия «2а» генерального корпуса современного монгольского языка, включающего разные жанры текстов: 1) художественная проза XX

века: романы; повести и рассказы; очерки; 2) поэзия XX в.; 3) перевод эпоса «Сокровенное сказание» на современный язык. 3) газетные тексты.

Корпус содержит 2394 текстов (длиной 2 126 510 слов).

Содержание корпуса выложено на сайте:
http://web-corpora.net/MongolianCorpus/search/index.php?interface_language=ru

Внесены уточнённые количественные данные в содержание подготавливаемой монографии «Монгольский язык в количественном аспекте: опыт корпусного исследования».

В 2012 году в состав корпуса включены газетные тексты из газеты “Даяар Монгол”. Их тематика разнообразна: среди текстов есть как посвящённые новостям по общим темам (таким, как администрация и управление, армия и вооруженные конфликты, астрология, парапсихология, эзотерика, бизнес, коммерция, экономика, финансы, дом и домашнее хозяйство, досуг, зрелища и развлечения, здоровье и медицина, искусство и культура, криминал, образование, политика и общественная жизнь, право, природа, производство, путешествия, религия, сельское хозяйство, спорт, строительство и архитектура, техника, философия, частная жизнь), так и научно-популярные: по естественным наукам (астрономии, биологии, географии, геологии, информатике, математике, статистике, физике, химии), по прикладным и технологическим областям (ветеринарии, военному делу, медицине, лесному хозяйству, производству, сельскому хозяйству, строительству и архитектуре, технике, транспорту, энергетике), по гуманитарным наукам (искусствоведению, истории, культурологии, лингвистике, образованию, политологии, праву, психологии, религиоведению, социологии, филологии, философии, экономике).

Включение в корпус многочисленных газетных текстов создаёт основания для количественного изучения лексических и грамматических характеристик газетных текстов. Квантитативный подход позволяет классифицировать сами газетные тексты в соответствии с языковыми стилями и жанрами, в рамках которых эти тексты создавались. Так как различия между этими стилями и жанрами носят преимущественно статистический характер, то таким образом

можно основать статистическую стилистику монгольского языка, описывающую и классифицирующую монгольские тексты по темам и жанрам на строго объективной базе.

Усовершенствован морфологический анализатор, уточнены словарь лексем и таблица омонимов. Проведена лемматизация и глоссирование корпуса в духе Лейпцигских правил глоссирования (с необходимой их адаптацией применительно к монгольскому языку), дальнейшая отладка анализатора.

Морфологический анализатор для современного монгольского языка написан на высокопроизводительном языке C, оформлен в виде независимой программной библиотеки и на данный момент работает под управлением информационной среды StarLing. Разработка находится на экспериментальной стадии, эффективно анализируется 96% текстовых словоформ (соответствующих 76% словоформ, являющихся входами в конкорданс словоформ, входящих в корпус).

Повышена эффективность программы частичного морфологического разбора (включающего отсылку либо к лексеме, либо к грамматеме).

Повышена эффективность программы полного морфологического разбора (включающего отсылку и к лексеме и к грамматеме).

Были созданы частотные словари монгольского языка

«Верхушки» некоторых ЧС можно привести уже сейчас. На материале прошлогодней версии (ГКМЯ-1) это уже проделано (в рабочем режиме).

Что касается столбца «количество текстов, в которых встретилась данная единица», который кажется довольно информативным, то здесь естественным образом возникает вопрос, почему в словарях словоформ и словарях лексем мы не видим никакого приближения к максимальной цифре (т. е. к единицам, встречающимся во всех текстах). Лишь показатель номинатива (NOM) в таблице грамматем более или менее приближается к этой цифре — 885. Причина такой странности заключена в следующем: в версии «ГКМЯ-1» 4% приходится на газетные статьи и 2% — на современную поэзию, где по понятным причинам представлены короткие и сверхкороткие тексты. В об-

щем случае число текстов, входящих в корпус, становится информативным, когда эти тексты примерно равны по объему (далеко не во всех корпусах соблюдается этот принцип); в противном случае параметр числа текстов отражает не столько употребительность единицы как таковую (во всем разнообразии текстов), сколько ее употребительность в составе тех языковых жанров, к которым принадлежат короткие и сверхкороткие (т. е. газетные и поэтические) тексты.

Частотность словоформ в монгольском языке.

Частотность лексем в монгольском языке

Так как работа велась по корпусу с неснятой омонимией, то иногда статус лексем получают не собственно лексемы, а дизъюнктивные пучки (частично) омонимических лексем. Однако информация о частотности таких единиц в корпусе является не менее ценной, нежели информация о частотности собственно лексем. Во всяком случае, для получения информации о количественных характеристиках лексики и грамматики МЯ не стоит ждать, пока будет создан корпус со снятой омонимией: ждать придется слишком долго, и в любом случае данные, полученные на основе корпуса со снятой омонимией, будут базироваться на малых эмпирических фактах, что обесценит их статистическую значимость.

Из «верхушек» языковых единиц в ранговых словарях можно извлечь немало ценных сведений, которые предоставляют почву для наблюдений типологического характера. Приведем несколько примеров таких наблюдений.

1) **Союз *бөгөөд* ‘и’** занимает всего лишь 52-е место в ЧС словоформ и 67-е место в ЧС лексем МЯ. Ср.: союз *and* в английских ЧС занимает 3-е (или 4-е, или 5-е) место; союз *и* в русском ЧС занимает 1-е место.

Такая большая разница в частотности союза ‘и’ в МЯ (с одной стороны), английском и русском языках (с другой) объясняется следующими факторами. Монгольский язык «предпочитает» бессоюзные конструкции: частота бессоюзия обычно компенсируется явлением т. н. алтайского типа сочине-

ния, предполагающего групповую флексию при сочинении существительных и богатство системы деепричастного таксиса в сфере глагола.

2) **Притяжательные местоимения** (особенно клитизованные) имеют очень высокую частоту в МЯ.

Например, ***нь*** (букв. **‘его, ее, их’**) занимает 1-е место в ЧС лексем и 1-е место в ЧС словоформ МЯ. Ср.: в английских ЧС *his* ‘его’ занимает 25-е (или 23-е, или 12-е) место, *her* ‘ее’ — 42-е (или 29-е, или 13-е) место, *its* ‘его’ — 78-е (или 77-е, или 142-е) место, *their* ‘их’ — 36-е (или 39-е, или 61-е) место. В русских ЧС притяжательное местоимение *его* занимает 41-е (или 50-е) место, *её* — 72-е (или 121-е) место, *их* — 86-е (или 134-е) место.

Притяжательное местоимение ***минь*** (букв. **‘мой’**) занимает 14-е место в ЧС словоформ и 20-е место в ЧС лексем МЯ. Ср.: в английских ЧС *my* ‘мой’ занимает 44-е (или 34-е, или 24-е) место; в русских ЧС *мой* занимает 60-е (или 69-е) место.

Притяжательное местоимение ***чинь*** (букв. **‘твой’**) занимает 18-е место в ЧС словоформ и 24-е место в ЧС лексем МЯ. Ср.: в английских ЧС *your* ‘твой’ занимает 69-е (или 64-е, или 62-е) место; в русских ЧС *твой* занимает 266-е (или 579-е) место.

Эта существенная разница объясняется тем, что в монгольском языке есть грамматическая категория притяжания, которая может быть выражена синтетически или аналитически. Аналитические средства выражения этой категории — энклитизованные притяжательные местоимения, в монгольском языке фактически играющие роль, сходную с той, которую во многих артиклевых европейских языках (например, романских и германских) играют артикли.

Если сравнить западноевропейские притяжательные местоимения с русскими, то можно убедиться, что западноевропейские притяжательные местоимения используются чаще, чем русские. Например, при переводе с русского языка на английский и обратно такие факты проявляются особенно ярко: ср. такие переводные эквиваленты, как *жена — my wife* (*your wife*, *his wife*), *мать — my mother* (*your mother*, *his mother*, *her mother*, *our mother*), *муж —*

my husband (*your husband*, *her husband*), *отец — my father* (*your father*, *his father*, *her father*, *our father*), *нос — my nose* (*your nose*, *his nose*, *her nose*), *голова — my head* (*your head*, *his head*, *her head*) и т. д.

Сравнение ЧС показывает, что монгольский язык, как и другие алтайские, продвинулся по шкале грамматикализации притяжательных местоимений еще дальше, чем западноевропейские языки. В результате этого продвижения сами притяжательные местоимения подверглись мощному процессу генерализации, транспозиции и десемантизации: они играют уже не только и не столько роль показателей притяжательности в собственном смысле слова, сколько роль показателей определенности, топикальности и субстантивации. Но такой процесс свойственен всем грамматическим категориям, так что сам факт наличия у притяжательных показателей вторичных (непритяжательных) функций на самом деле парадоксальным образом демонстрирует их грамматикализованный характер.

Частотность грамматем в монгольском языке

Так как работа велась по корпусу с неснятой омонимией, то иногда статус грамматем получают не собственно грамматемы, а дизъюнктивные пучки (частично) омонимических грамматем. Тем не менее, информация о частотности таких единиц в корпусе представляет не меньшую ценность, нежели информация о частотности собственно грамматем. Приведенные выше соображения о целесообразности использования корпуса с неснятой омонимией mutatis mutandis относятся и к грамматике.

Наличие в составе дизъюнктивных пучков грамматем таких пучков, в которых левый член дизъюнкции тождественен правому, объясняется тем, что в МЯ немало частично-омонимических пар, (а) внутри которых налицо отношение субкатегориальной конверсии (один из членов пары принадлежит к тематическому склонению, а другой — к атематическому), а также (б) члены которых различаются тем, что в исходе одного из них стоит «устойчивое» *н*, а в исходе другого — «неустойчивое» *н*.

Немало ценных сведений можно извлечь из анализа «верхушек» ранговых списков грамматем, такие сведения дают импульс для размышлений типологического характера. Остановимся для примера на следующих примечательных закономерностях.

1) **Деепричастные формы** глаголов (**конвербы**) занимают весьма высокие позиции в ЧС грамматем МЯ: конгрессивные деепричастия — 50195.83 ipm, модификативные деепричастия — 23820.50 ipm, антецессивные деепричастия — 8031.30 ipm.

В тех европейских языках, где есть деепричастия, они употребляются с гораздо более низкой частотностью; такая разница может быть объяснена следующим образом. Монгольский язык характеризуется т. н. алтайским типом сочинения, предполагающим богатую систему деепричастно выраженного таксиса в сфере глагола. Таким образом, он «предпочитает» бессоюзное сочинение: относительная редкость употребления союзов компенсируется богатством системы деепричастий.

2) **Возвратно-притяжательные формы** употребляются исключительно часто: «простые» возвратно-притяжательные формы — 10403.57 ipm, возвратно-притяжательные формы перфективных причастий — 2336.80 ipm, возвратно-притяжательные формы дательного падежа (датива) проспективных причастий — 2001.12 ipm и т. д.

Наиболее знакомые нам западноевропейские языки вообще не имеют такой грамматической формы. Русский язык, правда, имеет ближайший переводной эквивалент — возвратное местоимение *свой*. Однако сравнение относительной частотности местоимения *свой* с относительной частотностью возвратно-притяжательных форм в МЯ показывает, что *свой* употребляется с гораздо меньшей частотностью: его относительная частотность составляет 3825.5 ipm.

Это значительное количественное различие можно объяснить следующим образом: возвратно-притяжательные формы монгольских языков используются как одно из важнейших средств выражения кореферентности. Западно-

европейские артиклевые языки предпочитают иной способ выражения кореферентности: в этой функции используются преимущественно определенные артикли. Русский язык «предпочитает» т. н. «нулевые» формы выражения кореферентности: фактически они носят не столько «нулевой», сколько просодический (супрасегментный) характер; однако просодические механизмы не находят последовательного выражения в письменной форме языка (или, во всяком случае, находят его чрезвычайно редко).

6. Общее число опубликованных в 2012 г. по проекту работ
6.1. количество монографий 0
6.2. количество сборников статей 0
6.3. количество статей 2

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.)
Крылов С. А. Структурно-вероятностная модель монгольского языка на базе Генерального корпуса современного монгольского языка // Урало-алтайские исследования, № 1(6), с. 78-105.
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 версия «2а» генерального корпуса современного монгольского языка, включающего разные жанры текстов: 1) художественная проза XX века: 2) поэзия XX в.; 3) перевод эпоса «Сокровенное сказание» на современный язык. 3) газетные тексты. Корпус содержит 2394 текстов (длиной 2 126 510 слов).

Отличительные особенности версии «2а»- следующие.

Принципы лемматизации и глоссирования в ГКМЯ-2а. Объект лемматизации и глоссирования – микротакты. В качестве лемм используются: (а) синтетические лексемы; (б) лексикализованные сочетания синтетических лексем с клитиками (служебными словами). В качестве единиц грамматического глоссирования используются: (а) граммемы и сочетания граммем синтетически выражаемых грамматических категорий; (б) граммемы и сочетания граммем грамматических категорий, выражаемых как синтетическими средствами, так и клитиками (служебными словами).

Объём корпуса ГКМЯ-2а. Объём основной части корпуса ГКМЯ-2а: 2394 текста общим объёмом 2 126 510 словоупотреблений. Таким образом, за этот год корпус пополнился на 900 текстов длиной 971 485 словоупотреблений. Объём дополнительной части корпуса ГКМЯ-2а: 100 пар параллельных текстов (типа «оригинал-перевод») с объёмом монгольской части 129 тыс. словоупотреблений. Объём вспомогательной части корпуса ГКМЯ-2а: 10 текстов (размеченных со снятием омонимии) с объёмом монгольской части 11.5 тыс. словоупотреблений.

Содержание корпуса выложено на сайте:
http://web-corpora.net/MongolianCorpus/search/index.php?interface_language=ru

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

Создана версия «2а» генерального корпуса монгольского языка (ГКМЯ). ГКМЯ содержит 2394 текстов разных жанров (2 126 510 слов); т.о., за 2012 год ГКМЯ пополнился на 900 текстов (971 485 слов). Начата работа над корпусом русско-монгольских и монгольско-русских параллельных текстов, а также над корпусом со снятой омонимией. Морфологическая разметка охватывает не только синтетические, но и некоторые аналитические формы.

14. Краткий финансовый отчет за 2012 г. (основные статьи расходов по проекту, сумма) 280 000 руб.

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).

На втором этапе (2013 й год) предполагается создать версию 2б. Её отличительные особенности будут следующими.

1) Принципы лемматизации и глоссирования в ГКМЯ-2б. Объект лемматизации и глоссирования – мезотакты. В качестве лемм используются (а) синтетические лексемы; (б) лексикализованные сочетания синтетических лексем с клитиками (служебными словами); (в) лексикализованные сочетания синтетических лексем с полуслужебными (строевыми, делексикализованными) лексемами. В качестве единиц грамматического глоссирования используются: (а) граммемы и сочетания граммем синтетически выражаемых грамматических категорий; (б) граммемы и сочетания граммем грамматических категорий, выражаемых как синтетическими средствами, так и клитиками (служебными словами); (в) граммемы и сочетания граммем грамматических категорий, выражаемых как синтетическими средствами и клитиками, так и полуслужебными (делексикализованными) лексемами.

2) Объём корпуса ГКМЯ-2б.

2.1. Объём основной части корпуса ГКМЯ-2б: 3 тыс. текстов общим объёмом 3 млн. словоупотреблений.

2.2. Объём дополнительной части корпуса ГКМЯ-2б: 200 пар параллельных текстов (типа «оригинал-перевод») с объёмом монгольской части 200 тыс. словоупотреблений.

2.3. Объём вспомогательной части корпуса ГКМЯ-2б: 20 текстов (размеченных со снятием омонимии) с объёмом монгольской части 20 тыс. словоупотреблений.

Подпись руководителя проекта С.А. Крылов

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
	«Генеральный корпус совре-менного монгольского языка, версии 2а-2в»	ИВ РАН	Крылов С. А. (+ Пюрбеев Г.Ц., Яхонтова Н.С., Петрова М.П.)		2013 г. — создание версии ГКМЯ-2б (син-тетическая и аналитическая морфология, грамматикализованные сочетания; 3 тыс. текстов на 3 млн. словоупотреблений; 200 параллельных текстов; 20 текстов со снятой омонимией)