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 Корпусная лингвистика. Направление 1. Создание и развитие корпусных ресурсов по современному русскому языку.</td></tr>
<tr><td colspan="2">Название проекта
Синтаксическая разметка корпуса со снятой лексико-грамматической омонимией НКРЯ</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Ляшевская Ольга Николаевна, к.ф.н., с.н.с. ИРЯ им. В.В.Виноградова РАН</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)
почтовый адрес руководителя проекта</td><td>olesar@gmail.com</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования
Учреждение Российской академии наук Институт русского языка им. В. В. Виноградова РАН (ИРЯ РАН)</td><td>ФИО (полностью) руководителя организации – адресата финансирования директор Института русского языка им. В.В. Виноградова РАН (ИРЯ РАН) чл.-корр. РАН 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>Толдова Светлана Юрьевна, к.ф.н., Российский государственный гуманитарный университет</td></tr>
<tr><td>Соколова Елена Григорьевна, к.ф.н., Российский государственный гуманитарный университет</td></tr>
<tr><td>Сичинава Дмитрий Владимирович, к.ф.н., ИРЯ РАН
Летучий Александр Борисович, к.ф.н., ИРЯ РАН
Бонч-Осмоловская А.А., к.ф.н., НИУ ВШЭ
Привознов Д.К., МГУ
Ионов М.К., МГУ
Гарейшина А.Г., МГУ</td></tr>
<tr><td>Дата сдачи отчета 20/11/2012</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

   Синтаксическая разметка корпуса со снятой лексико-грамматической омонимией НКРЯ.

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

   Ляшевская Ольга Николаевна, к.ф.н., с.н.с. ИРЯ им. В.В.Виноградова РАН

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

   Толдова Светлана Юрьевна, к.ф.н., Российский государственный гуманитарный университет;

   Соколова Елена Григорьевна, к.ф.н., Российский государственный гуманитарный университет

   Сичинава Дмитрий Владимирович, к.ф.н., н. с. ИРЯ РАН

   Летучий Александр Борисович, к.ф.н., н. с. ИРЯ РАН

   Бонч-Осмоловская Анастасия Александровна, к.ф.н., профессор НИУ ВШЭ

   Привознов Дмитрий Константинович, студент МГУ

   Ионов Максим, студент МГУ

   Гарейшина Анастасия, студент МГУ

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

Общей задачей проекта является разработка методов интеграции частичной синтаксической разметки в НКРЯ со снятой лексико-грамматической

омонимией, создание Золотого стандарта: подкорпуса объёмом в 500 тыс. словоупотреблений с ручной разметкой наиболее важных типов синтаксической связи.

В рамках проекта в 2012 году были запланированы следующие виды работ: разработка технологии разметки, основанной на сравнении результатов синтаксической аннотации корпуса несколькими автоматическими синтаксическими анализаторами, разработка принципов аннотации, разметка корпуса объемом 100 тыс. словоупотреблений по нескольким основным типам синтаксическим типам связи, разработка поискового интерфейса к корпусу с синтаксической разметкой.

Задачи первого этапа проекта были выполнены. Для разметки был выбран и подготовлен корпус объемом 100 тыс. словоупотреблений (порядка 6 тысяч предложений). Материалом корпуса послужил фрагмент случайной выборки предложений, предоставленных для свободного пользования на сайте Национального корпуса русского языка (URL: http://ruscorpora.ru/corpora-usage.html). Корпус с ручной разметкой базовых синтаксических отношений доступен в Интернете по адресу http://rus-treebank.soiza.com/. По результатам сравнения синтаксических разборов корпуса текстов (1 млн. словоупотреблений) различными синтаксическими парсерами была разработана основная технология синтаксической разметки в полуавтоматическом режиме с ручной проверкой. Было решено осуществлять разметку корпуса по конструкциям (отдельным типам синтаксических связей). На первом этапе в аннотации корпуса были учтены следующие базовые типы синтаксической связи: ‘предикат - подлежащее’, ‘предикат – прямое дополнение’, ‘существительное - согласованное определение’. Данные типы отношений были отобраны для разметки в первую очередь. Они были исправлены автоматически, после чего вручную проверены аннотаторами. Был также отобран ряд синтаксических связей и конструкций, которые будут интегрированы в синтаксическую разметку корпуса во вторую очередь. Они были детально проанализированы. Были исчислены случаи расхождений в ответах разных систем, типичные ошибки анализа, спорные и омонимичные случаи. Были предложены автоматические методы первичной коррекции результатов анализа этих конструкций системой SyntAutom для последующей ручной проверки данных типов отношений и интеграции результатов коррекции в корпус в ближайшие два месяца. Результаты обобщений относительно типов вариативности синтаксического разбора, относительно типичных ошибок автоматического синтаксического анализа вошли в доклад на международной конференции «Диалог», посвященный оценкам методов автоматического анализа текстов в области синтаксического анализа, также они вошли в постерный доклад, который должен быть представлен на международной конференции Coling-2012 в начале декабря 2012г. Текст статьи опубликован в сборнике трудов конференции в рамках, а также в материалах конференции Coling-2012.

Работа над проектом включала следующие этапы:

1) Предварительная подготовка корпуса

2) Формирование банка расхождений синтаксических аннотаций, полученных разными системами автоматического синтаксического анализа и выбор базового набора синтаксических отношений для первого этапа синтаксической разметки корпуса;
3) Разработка технологии разметки корпуса в полуавтоматическом режиме;
4) Создание редактора для разметки корпуса;
5) Разметка корпуса объемом 100 тыс. словоупотреблений.

Уточнение задачи

Основными задачами создания синтаксически размеченного корпуса в рамках настоящего проекта являются: разработка принципов интеграции синтаксической разметки в Национальный корпус русского языка, а также создание корпуса для обучения так называемых «легких» синтаксических парсеров. Последние используются в задачах автоматического извлечения информации из текстов и, как правило, требуют достаточно грубого синтаксического анализа, при котором надежного должно распознаваться только небольшое количество основных синтаксических связей. Обе данные задачи накладывают некоторые ограничения на принципы синтаксической разметки, а именно, число "различаемых" синтаксических связей должно быть не очень большим. Выделяемые связи должны хорошо соотноситься с представлением неподготовленного пользователя о синтаксической структуре предложения, то есть, по возможности, быть достаточно теоретически нейтральны. На предварительном этапе разработки основных принципов синтаксической разметки корпуса в рамках данного проекта был проведен сравнительный анализ существующих систем автоматического синтаксического анализа для русского языка, а также анализ организации и принципов разметки аналогичных существующих корпусов. В частности в качестве образцов рассматривался синтаксический корпус СинТагРус, созданный в Лаборатории компьютерной лингвистики Института проблем передачи информации РАН, а также синтаксический корпус чешского языка The Prague Dependency Treebank. Также был учтен опыт ручной синтаксической разметки на основе инструкции Е.Г.Соколовой (доступна с сайта http://testsynt.soiza.com/files/info.htm).

В силу вышеуказанных причин в настоящем корпусе была принята упрощенная схема синтаксической разметки. В то же время по возможности целый ряд решений, принятых в СинТагРус, были взяты за образец в целом ряде спорных случаев.

Представление данных корпуса

Корпус представляет собой базу данных, доступ к которой осуществляется через специально разработанный Web-интерфейс. Каждому словоупотреблению (токену) приписана ее морфологическая характеристика, а также номер ее синтаксического «хозяина» и тип связи зависимого с главным (например, в предложении *(1) Каких (2) именно (3) результатов (4) можно (5) ждать* – синтаксическая информация относительно связи *ждать -> результатов* отражена в базе следующим образом: 3. результатов; head: 5; type: obj)

Основные этапы и результаты по задачам, решаемым в рамках проекта в 2012 г.

Этап 1. Предварительная обработка корпуса

В связи с тем, что технология синтаксической разметки в рамках данного проекта, по крайней мере, на начальном этапе предполагала сравнение результатов анализа работы разных систем и выявление различных типов расхождений, необходимо было провести предварительную подготовку корпуса. При создании корпуса с параллельной синтаксической разметкой в рамках проекта корпусной программы РАН «Совершенствование разметки и системы выдачи данных в Национальном корпусе русского языка» для анализа системам был предложен корпус с предварительной разбиением на предложения и токены. Однако не все системы соблюдали это разбиение. Возник целый ряд проблем с разбиением на токены сложных лексических единиц, таких как слова с дефисом (например, *серо-голубой*) или сложных союзов или предлогов (например, в течение, *при помощи* и т.п.). При этом, если для сравнения результатов морфологического анализа нарушение нумерации токенов может быть легко скорректировано, то в синтаксическом анализе такое нарушение вдет к тому, что целый фрагмент синтаксического дерева имеет ошибочную нумерацию вершин. Таким образом, на этапе подготовки корпуса был разработан модуль автоматического выравнивания, осуществляющий исправление ошибок парсеров в разбиении исходного корпуса на предложения и унификацию расхождений относительно анализа сложных лексических единиц, было произведено выравнивание результатов работы разных парсеров по токенам, а также разработана система возможной ручной правки результатов выравнивания с пересчетом всех номеров токенов и их синтаксических «хозяев».

Этап 2. Выбор базового синтаксического разметчика. Выбор типов синтаксической связи для первого этапа синтакасической разметки корпуса

2.1. Выбор разметчика и сравнение результатов автоматического анализа, представленного разными системами. Анализ расхождений по отдельным типам связей

В качестве базового синтаксического разметчика был выбран SyntAutom, разработчики которого были готовы содействовать в дальнейшей разметке корпуса, а также приняли участие в обсуждении основных принципов разметки.

На первом этапе было произведено сравнение результатов работы данного синтаксического анализатора и других анализаторов, представленных в корпусе параллельной синтаксической разметки (http://testsynt.soiza.com/). Были также учтен опыт проведения Форума по оценке методов автоматического анализа языка.

В результате был выявлен целый ряд конструкций, для которых свойственна наибольшая вариативность в ответах систем относительно направлений синтаксической связи, был создан реестр допустимых общетеоретических расхождений и таблицы их «эквивалентности» (таблица доступна по адресу: http://testsynt.soiza.com/files/var-synt.htm).

Для выделения типов синтаксических связей, в первую очередь подлежащих разметке, исследовалась частотность конструкций с различным типом синтаксической связи, были выявлены менее «устойчивые» и более «устойчивые» к ошибкам типы синтаксической связи.

Так, например, одним из наиболее проблемных мест в пределах простого предложения оказались синтаксические связи, с предлогом в качестве зависимого. Этот тип связи представляет собой случай один из наиболее частотных случаев синтаксической омонимии, а также один из случаев, неустойчивых к ошибкам анализа. Например, в примере (1) 7 из автоматических систем сделали ошибку, только один разбор оказался правильным:

(1) Сургут ***уступил*** *семь* ***<u>баллов</u> по показателям*** *износа теплосетей, канализации, систем водоснабжения.*

- *(1) – баллов* ***-->*** *по (7 систем)*
  (2) – уступил ***-->*** *по (1 система)*

Сравнительный анализ систем показал, что разные системы не только используют разные названия для одних и тех же синтаксических отношений, но существуют значительные расхождения в самой классификации типов связи. Например, в одних системах разграничение типов связей опирается на морфологическую разметку, в других, наоборот, учитывается самая общая синтаксическая функция словоформы.

Для задач данного корпуса было решено опираться в большей степени на общие синтаксические функции словоформ. Однако для удобства поиска по синтаксически размеченному корпусу нестандартных синтаксических связей, синтаксических связей, отличающихся от некоторого прототипа, например, так называемых дативных подлежащих, а также для возможности различать ситуации, когда одно обобщенное синтаксическое отношение приписывается очень разным по своим морфологическим и частеречным свойствам зависимым, было решено в качестве отдельного параметра ввести дополнительную детализацию разметки. Так, например, различать отношение типа ‘предикат – каноническое подлежащее’ – Subj и ‘предикат – неканоническое подлежащее’ – Subj:NonNom.

В результате детального анализа вариативности и ошибок систем в различных классах синтаксических конструкций различные типы синтаксических конструкций были разбиты на несколько классов: (а) наиболее частотные и значимые синтаксические связи, которые должны быть отражены в корпусе в первую очередь, такие как , таких как ‘подлежащее – предикат’, ‘предикат – прямое дополнение’, согласованные определения; (б) значимые синтаксические связи, вызывающие проблемы при синтаксическом анализе: вариативность и/или ошибки при анализе разными системами, которые требуют более детальной проработки и должны быть интегрированы во вторую очередь, такие как, например, синтаксические связи, в которых участвуют предлоги, инфинитивные обороты, конструкции с именными предикатами и др.; (в) остальные.

Таким образом, было проведено сравнение работы синтаксических анализаторов по целому ряду синтаксических конструкций, внутри каждой из

которых были выделены различные случаи теоретического варьирования, типовых ошибок, нестандартных решений. Также все случаи были расклассифицированы по следующим параметрам: случаи, когда возможен автоматический пересчет решений одной системы относительно другой; регулярные ошибки одной из систем, которые можно исправить, опираясь на анализ работы другой системы; случаи, требующие систематического просмотра; сложные случаи синтаксической омонимии, вызывающие расхождения в разборе при ручной разметке; случайные ошибки системы, связанные, например, с морфологической омонимией.

2.2. Создание редактора исправления ошибок

Как отмечалось выше, задача ручной разметки на первом этапе предполагала предварительную автоматическую коррекцию ошибок анализа, выполненного системой SynAutom, а также «пересчет» решений SynAutom в соответствии с разработанными принципами разметки. В ходе решения данной задачи была разработана система автоматического пересчета связей. Для редактирования результатов автоматической коррекции был усовершенствован специальный редактор, созданный на базе разработанной в рамках проекта 2010-2011 гг. системы визуализации. Данная система сравнивает результаты разметки, визуализирует результаты сравнения, автоматически выделяя проблемные места разметки – места несовпадений результатов анализа. Как показывает анализ работы данной системы, такие места несовпадений позволяют выявить и типизировать ошибки автоматического анализатора.

2.3. Усовершенствование системы поиска

Пилотная версия поиска по синтаксическим связям была разработана в рамках Корпусной программы 2010-2011 гг. Также было разработано два возможных варианта результатов поиска: в виде деревьев и в табличном виде. Тестирование этих вариантов представления данных показало, что оба формата являются востребованными. В систему поиска была добавлена возможность использовать в качестве условий поиска морфологические характеристики. Была добавлена возможность выгрузки некоторого количества примеров из выдачи в таблице Excel.

Этап 3. Разметка корпуса по отдельным типам связей

Как указывалось выше, по результатам этапа 2 были выделены целевые типы связи для отработки методики полуавтоматической разметки и для разметки первой части корпуса, была разработана сама методика создания корпуса с синтаксической разметкой. В результате был размечен корпус объемом 100 тыс. словоупотреблений по базовым типам синтаксических связей.

Публикации по проекту:

Толдова С.Ю., Соколова Е.Г., Астафьева И., Гарейшина А., Королева А., Привознов Д., Сидорова Е., Тупикина Л., Ляшевская О.Н. Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка // *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»*. Вып. 11(18). М.: Изд-во РГГУ. С. 797-809.

Anastasia Gareyshina, Maxim Ionov, Olga Lyashevskaya, Dmitry Privoznov, Elena Sokolova and Svetlana Toldova. RU-EVAL-2012: Evaluating dependency parsers for Russian // *COLING-2012* (в печати).

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

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)

Подготовлен и размещен в интернете корпус с синтаксической разметкой объемом 100 тыс. словоупотреблений. На данном этапе в корпусе отражены такие базовые типы синтаксической связи как 'предикат - подлежащее', 'предикат – прямое дополнение', 'существительное - согласованное определение'. Текстовым материалом для разметки послужил фрагмент случайной выборки предложений объемом порядка 6 тыс. предложений, выложенный для свободного пользования на сайте Национального корпуса русского языка. Разметка производилась на базе первичной автоматической синтаксической разметки. В процессе работы над корпусом была разработана методика частичной автоматизации ручной разметки с использованием результатов параллельной синтаксической разметки предложений несколькими парсерами, а также система редактирования связей. В процессе создания корпуса были решены следующие задачи: оценка «устойчивости» к ошибкам разных типов синтаксических конструкций, определение наиболее проблемных мест для автоматических синтаксических анализаторов, анализ расхождений работы различных синтаксических парсеров, представленных в тестовом корпусе с параллельной синтаксической разметкой. По итогам работы опубликованы статьи.

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

Подготовлен и размещен в интернете корпус с синтаксической разметкой, проверенной вручную, объемом 100 тыс. словоупотреблений. В корпусе отражены базовые типы синтаксической связи: 'предикат - подлежащее', 'предикат – прямое дополнение', 'существительное - согласованное определение'. Разработана общая технология синтаксической разметки для дальнейшей разметки корпуса с учетом других типов синтаксической связи.

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).

Для выполнения общей задачи проекта необходимо на втором этапе расширить список синтаксических связей, по которым размечается создаваемый корпус. Также необходимо пополнить корпус размеченными по данным типам синтаксических связей текстами, общим объемом 200 тыс. словоупотреблений.

В результате выполнения первого этапа сформировано два множества синтаксических отношений. Для первого множества уже существует база сравнительного анализа разборов этих отношений разными системами, в которой исчислены разные типы ситуаций, различающихся разными типами варьирования, а также разными типами стандартных ошибок. Для данного множества отношений задача будет состоять в разработке стандартов разметки: ярлык для имен синтаксических отношений, выработка окончательного решения относительно направлений связи, а также в разработке методов полуавтоматической коррекции исходной автоматической разметки и методов проверки результатов ручной коррекции. Это, в частности такие отношения как внешние и внутренние связи с предлогом, связи в инфинитивных конструкциях, связи конструкциях со связочными глаголами и с нулевыми связками. Для второй группы отношений необходимо будет вначале создать базу расхождений и ошибок. Это такие отношения как связи предиката с сирконстантами, связи в именной группе с несогласованными определениями, связи между частями сложных предложений и др. При работе над каждой из типов конструкций будут учитываться частотные характеристики данных типов конструкций.

Дальнейшая разметка корпуса, пополнение корпуса до 300 тыс. словоупотреблений будет происходить на основе соответствующей работы над двумя перечисленными выше группами синтаксических связей.

Также предполагается активное тестирование системы поиска по синтаксическому поиску и дальнейшее развитие поискового интерфейса.

Подпись руководителя проекта

Форма 2
Планируемое содержание работ на 2013 г.

№	Название проекта	Организация-исполнитель и учреждения-соисполнтели	Руководитель проекта (+ кол-во исполнителей)	Запрашиваемый объем финансирования на 2013 г. (тыс. руб.)	Ожидаемые в 2013 г. результаты
	Синтаксическая разметка корпуса со снятой лексико-грамматической омонимией НКРЯ	Института русского языка им. В.В. Виноградова РАН (ИРЯ РАН)	Ляшевская Ольга Николаевна, к.ф.н., с.н.с. (+ 2 исполнителя)		Пополнение синтаксически размеченного корпуса до 300 тыс. словоупотреблений, разметка его по всем основным типам синтаксических связей.