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
Разработка системы морфологического и синтаксического анализа русских текстов на основе корпуса СинТагРус</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)
Богуславский Игорь Михайлович, доктор филологических наук, профессор</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города), почтовый адрес руководителя проекта</td><td>e-mail: bogus@iitp.ru; тел. (495)699-49-27; факс (495)650-05-79, 127994, Москва, ГСП-4, Большой Каретный пер. 19, стр. 1, ИППИ РАН</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования

Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН)</td><td>ФИО (полностью) руководителя организации – адресата финансирования
Кулешов Александр Петрович, академик РАН</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования
Лобачёва Римма Ивановна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования
127994, Москва, ГСП-4, Б.Каретный пер., д. 19, стр. 1, ИППИ РАН; тел. (495)650-42-25; факс (495)650-05-79;
e-mail: director@iitp.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="5">Исполнители (ФИО, уч. ст.)</td><td>Диконов Вячеслав Григорьевич</td></tr>
<tr><td>Дяченко Павел Владимирович, к.т.н.</td></tr>
<tr><td>Иомдин Леонид Лейбович, к.филол.н.</td></tr>
<tr><td>Петроченков Вадим Викторович</td></tr>
<tr><td>Цинман Леонид Львович, к.ф.-м.н.</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. РАН 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

Разработка системы морфологического и синтаксического анализа русских текстов на основе корпуса СинТагРус.

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

Богуславский Игорь Михайлович, доктор филологических наук, заведующий лабораторией, ИППИ им. А.А. Харкевича РАН

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

Диконов Вячеслав Григорьевич, младший научный сотрудник, ИППИ им. А.А. Харкевича РАН;

Дяченко Павел Владимирович, кандидат технических наук, научный сотрудник, ИППИ им. А.А. Харкевича РАН;

Иомдин Леонид Лейбович, кандидат филологических наук, ведущий научный сотрудник, ИППИ им. А.А. Харкевича РАН;

Петроченков Вадим Викторович, стажер-исследователь, ИППИ им. А.А. Харкевича РАН;

Цинман Леонид Львович, кандидат физико-математических наук, ведущий научный сотрудник, ИППИ им. А.А. Харкевича РАН.

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

В 2012 году работа по проекту была сосредоточена на двух направлениях: во-первых — на морфологическом теггере, а во вторых — на модернизации существующей программной части морфо-синтаксического анализатора системы ЭТАП-3.

1) Морфологический теггер.

В 2011 году на основе системы SVMTool (http://www.lsi.upc.edu/~nlp/SVMTool/) и синтаксически размеченного корпуса русских текстов СинТагРус был получен морфологический теггер для русского языка, показавший достаточно хорошие результаты по разметке (94% корректно расставленных тегов).

Однако возможности для настройки SVMTool ограничены, и некоторые идеи, способные улучшить качество разметки, не могут быть реализованы при использовании SVMTool «как есть». Тремя наиболее существенными недостатками этого тег-

гера являются: 1) невозможность работы с тегами как сложными объектами, имеющими подтеги (что существенно для морфологической разметки русского языка); 2) использование SVMLight в качестве единственной доступной библиотеки, реализующей метод опорных векторов; 3) невозможность сопряжения с существующим морфологическим анализатором системы ЭТАП, особенно при разметке текстов корпуса.

Поэтому было принято решение о собственной реализации части функционала, доступного в SVMTool, которая была расширена, чтобы отвечать, в частности, требованиям по отсутствию трёх вышеуказанных недостатков. Эта работа была проведена в 2012 году.

При разметке данный теггер, как и SVMTool, проходит по предложению и для каждого слова с помощью метода опорных векторов выбирает корректный тег на основе данных о самом слове и его контексте. При этом используется расширенная модель данных о контексте, включающая информацию о подтегах, и специализированная модель для незнакомых слов. В качестве библиотеки, реализующей метод опорных векторов, используется liblinear (http://www.csie.ntu.edu.tw/~cjlin/liblinear/), которая лучше соответствует специфике задачи.

Входными данными при обучении и разметке являются выходные данные морфологического анализатора системы ЭТАП-3, а именно — разбиение на слова и список возможных тегов для всех знакомых слов. Выходными данными является набор вероятностей для каждого возможного тега и, соответственно, наиболее вероятный тег. Пример результата работы морфологического анализатора и теггера (в виде возможных разборов, их тегов и вероятностей) можно увидеть на рис. 1.

Полученный теггер при разметке обрабатывает порядка 1500 слов в секунду и обеспечивает качество разметки (включая теги знаков препинания): для неоднозначных слов — 92.6%, для незнакомых слов — 73.0%, для всех слов — 96.0%.

Теггер был встроен как составная часть в общий анализатор системы ЭТАП и прорабатывает между морфологическим и синтаксическим анализатором. Синтаксический анализатор затем пользуется выходными данными теггера, что позволяет ему улучшить качество получаемой итоговой разметки и скорость работы. Качество морфологической разметки в такой связке достигает 96.2% (не включая теги знаков препинания). Встраивание этой компоненты служит первой частью превращения ЭТАПа в гибридный правилово-статистический парсер, второй частью послужит встраивание аналогичной компоненты для синтаксической разметки.

```
1.1  ЭТОТ                      A, ИМ, МН, CAPIT, CAP (p=0.941154)
1.2  ЭТОТ                      A, ВИН, МН, НЕОД, CAPIT, CAP (p=0.0588465)
2.1  ТИП1                      S, ИМ, МН, МУЖ, НЕОД (p=0.758434)
2.2  ТИП1                      S, ВИН, МН, МУЖ, НЕОД (p=0.00480953)
2.3  ТИП2                      S, ИМ, МН, МУЖ, ОД (p=0.236757)
3.1  СТАЛЬ                     S, РОД, ЕД, ЖЕН, НЕОД (p=0.000544517)
3.2  СТАЛЬ                     S, ДАТ, ЕД, ЖЕН, НЕОД (p=9.41069e-005)
3.3  СТАЛЬ                     S, ПР, ЕД, ЖЕН, НЕОД (p=5.98699e-005)
3.4  СТАЛЬ                     S, ИМ, МН, ЖЕН, НЕОД (p=0.000414142)
3.5  СТАЛЬ                     S, ВИН, МН, ЖЕН, НЕОД (p=9.43599e-005)
3.6  СТАНОВИТЬСЯ1              V, ПРОШ, МН, ИЗЪЯВ, СОВ (p=0.332931)
3.7  СТАНОВИТЬСЯ2              V, ПРОШ, МН, ИЗЪЯВ, СОВ (p=0.332931)
3.8  СТАТЬ1                    V, ПРОШ, МН, ИЗЪЯВ, СОВ (p=0.332931)
4.1  БЫТЬ                      V, НАСТ, ЕД, ИЗЪЯВ, 1-Л, НЕСОВ (p=0.034161)
4.2  БЫТЬ                      V, НАСТ, ЕД, ИЗЪЯВ, 2-Л, НЕСОВ (p=0.0316204)
4.3  БЫТЬ                      V, НАСТ, ЕД, ИЗЪЯВ, 3-Л, НЕСОВ (p=0.196774)
4.4  БЫТЬ                      V, НАСТ, МН, ИЗЪЯВ, 1-Л, НЕСОВ (p=0.0853929)
4.5  БЫТЬ                      V, НАСТ, МН, ИЗЪЯВ, 2-Л, НЕСОВ (p=0.079042)
4.6  БЫТЬ                      V, НАСТ, МН, ИЗЪЯВ, 3-Л, НЕСОВ (p=0.49188)
4.7  ЕСТЬ1                     V, ИНФ, НЕСОВ (p=0.0786441)
4.8  ЕСТЬ2                     INTJ (p=0.00248547)
5.1  ФИКТ-КОМПОЗИТ(В-ИНИЦИАЛ)  COM, CAP-MIX, STRICT_ABBR (p=0.0417274)
5.2  В1                        PR (p=0.319424)
5.3  В2                        PR (p=0.319424)
5.4  В3                        PR (p=0.319424)
5.5  ВЕК                       S, ИМ, ЕД, МУЖ, НЕОД, TRUN, STRICT_ABBR (p=8.19634e-007)
5.6  ВЕК                       S, РОД, ЕД, МУЖ, НЕОД, TRUN, STRICT_ABBR (p=7.6757e-008)
5.7  ВЕК                       S, ДАТ, ЕД, МУЖ, НЕОД, TRUN, STRICT_ABBR (p=1.60943e-007)
5.8  ВЕК                       S, ВИН, ЕД, МУЖ, НЕОД, TRUN, STRICT_ABBR (p=2.72816e-008)
5.9  ВЕК                       S, ТВОР, ЕД, МУЖ, НЕОД, TRUN, STRICT_ABBR (p=1.22801e-007)
5.10 ВЕК                       S, ПР, ЕД, МУЖ, НЕОД, TRUN, STRICT_ABBR (p=7.85442e-008)
6.1  ЦЕХ1. [ТЧК]               S, ПР, ЕД, МУЖ, НЕОД (p=0.5)
6.2  ЦЕХ2. [ТЧК]               S, ПР, ЕД, МУЖ, НЕОД (p=0.5)
```

Рис. 1.

2) Модернизация программной части синтаксического анализатора ЭТАП-3

Правиловая часть анализатора ЭТАП-3, использующая при работе накопленную за долгое время базу лингвистических знаний, состоящую из правил и словарей, остаётся основой синтаксического анализа в системе ЭТАП-3 и залогом его высокого качества.

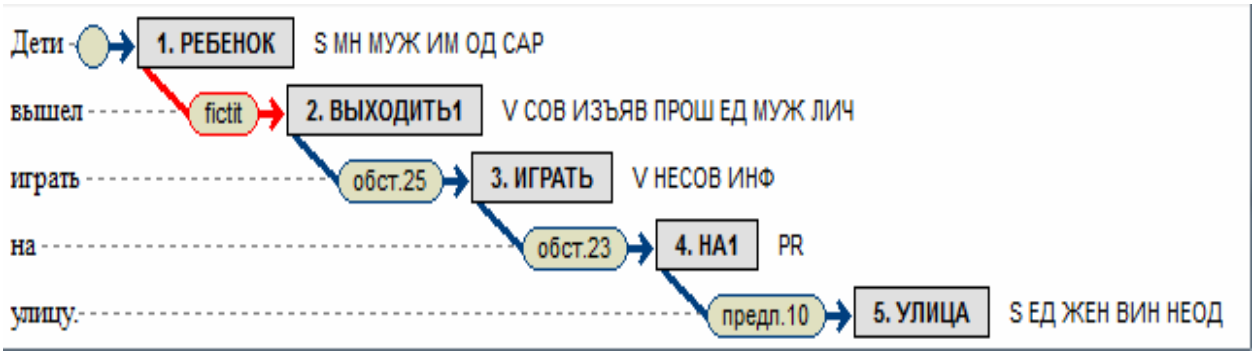
Разные части программного комплекса, реализующего преобразование этой базы знаний в действующий алгоритм синтаксического анализа были написаны в разное время и в разных парадигмах программирования. Некоторые проектные решения, принятые при его разработке, устарели, некоторые были неудачными, а некоторые ограничивают переносимость системы ЭТАП на другие платформы. Реализация некоторых частей алгоритма далека от оптимальной. Поэтому считается целесообразной и ведётся работа по модернизации и оптимизации кода существующего правилового парсера.

Видимыми внешнему пользователю результатами работ, проведённых за отчётный период, является исправление серии ошибок в работе парсера, расширение возможностей языка описания правил и увеличение скорости работы морфосинтаксического анализатора более чем в 2 раза. В среднем, теперь синтаксический анализатор обрабатывает около 6-7 предложений в секунду, что позволяет расширить область его возможных практических применений.

Также в алгоритм синтаксического анализа были внесены изменения, направленные на стабилизацию его работы с «плохими» предложениями, содержащими

ошибки и рассогласования, для которых система правил не могла получить приемлемую структуру. Искажения синтаксической структуры, вносимые подобными ошибками сделаны по возможности наиболее локальными и не портящими структуру в целом. Данное изменение призвано повысить применимость анализатора к реальным текстам.

В качестве примера рассмотрим разбор предложения с ошибкой согласования «Дети вышел играть на улицу». Хотя правильная предикативная связь между ВЫХОДИТЬ и РЕБЕНОК не строится, она локально заменяется фиктивной связью.



Кроме того, определённые изменения в правиловом анализаторе требуются как предварительный шаг перед введением в него полноценной статистической компоненты.

На настоящий момент качество работы синтаксического анализатора составляет около 89.5% правильно угаданных неименованных синтаксических связей и 86.4% правильно угаданных именованных синтаксических связей, что находится на уровне лучших результатов, полученных для русского языка.

6. Общее число опубликованных в 2012 г. по проекту работ

6.1. количество монографий

6.2. количество сборников статей

6.3. количество статей

7. Список опубликованных монографий и сборников статей, с полным указанием выходных данных, объема /в п.л. и количество стр./, а также их краткие аннотации (до 0,5 стр.)

8. Список опубликованных по проекту статей (объемом не менее 1 п.л.)

1. Leonid Iomdin, Vadim Petrochenkov, Victor Sizov, Leonid Tsinman. ETAP parser: state of the art. (Л.Л.Иомдин, В.В.Петроченков, В.Г. Сизов, Л.Л.Цинман Синтаксический анализатор системы ЭТАП: современное состояние). // Dialog 2012. Computational Linguistics and Intellectual Technologies. (International Conference. Moscow, RGGU Publishers, Issue 11(18). P. 830-843 (на англ. языке)

2. Leonid Iomdin. Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact. Jazykovedné štúdie, Bratislava (на англ. языке, в печати).

3. В.В.Петроченков. Морфологическая разметка русскоязычных текстов с помощью теггера на основе SVM. http://itas2012.iitp.ru/pdf/1569601641.pdf

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)

Был построен морфологический теггер для русского языка, объединяющий как статистический (на основе корпуса СинТагРус), так и словарный (на основе словаря и морфологического анализатора системы ЭТАП-3) подходы. Результатом работы теггера является полная морфологическая разметка (части речи и прочие характеристики — падеж, число, род и т. д.) в формате, принятом в корпусе СинТагРус. Процент правильно расставленных тегов составляет: для неоднозначных слов — 92.6%, для незнакомых слов — 73.0%, для всех слов — 96.0%. Данные показатели находятся на уровне лучших результатов, полученных для русского языка.

Результатом работы теггера является не только верный тег, но и вероятности для всех возможных тегов, что позволило встроить его в синтаксический анализатор системы ЭТАП-3 в качестве компонента. Данный компонент позволил улучшить как качество синтаксического анализа, так и его скорость за счёт раннего отсева маловероятных омонимов.

Улучшена стабильность работы правилового синтаксического парсера в случае «плохих» предложений, когда система правил не может выделить приемлемую синтаксическую структуру. К таким случаям зачастую относятся и предложения с ошибками согласования, опечатками и подобными случаями, которые теперь могут получить адекватный разбор.

Была продолжена работа по общей модернизации и оптимизации программной части анализатора системы ЭТАП-3, которая привела к увеличению стабильности его

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

На основе корпуса СинТагРус был построен высокоточный статистический морфологический теггер для русского языка. Теггер был включён в состав морфосинтаксического анализатора системы ЭТАП-3 и улучшил показатели качества и скорости анализа. Частично модернизирована и оптимизирована программная часть правилового синтаксического анализатора, повышена скорость, стабильность и качество анализа.

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 предполагается провести следующую работу:

- Усовершенствовать метод машинного обучения SVM на основе новейших разработок (Reverse Kernel Engineering, Pighin and Moschitti, 2009). Этот подход позволяет реализовать более эффективный алгоритм выделения релевантных признаков (фрагментов деревьев) в пространстве древесных ядер.
- Разработать модуль разрешения синтаксических неоднозначностей на основе SVM. Наличие таких неоднозначностей создает значительные трудности при

выборе правильной структуры, и мы ожидаем, что метод опорных векторов позволит охватить такие типы неоднозначностей, которые до сих пор не поддавались разрешению.

- Будет построен действующий прототип гибридного парсера, использующий полный набор синтаксических правил системы ЭТАП-3 и машинное обучение. Это потребует пересмотра и уточнения большого числа правил и новой привязки их к словарю системы.

- Будет продолжена работа по модернизации программного комплекса системы ЭТАП-3.

- Предполагается написание двух-трех научных статей и выступление с докладами, на конференциях по корпусной и компьютерной лингвистике.

Руководитель проекта
профессор Богуславский И.М.

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
36П	Разработка системы морфологического и синтаксического анализа русских текстов на основе корпуса СинТагРус	ИППИ РАН	Богуславский И.М. + 5		Эксперименты с гибридным парсером на материале корпуса СинТагРус, на материале части Национального корпуса русского языка со снятой омонимией. Развитие гибридного парсера. Разработка модуля разрешения неоднозначностей на основе SVM.