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<br>Направление 4. Создание и развитие корпусных ресурсов по языкам мира.</td></tr>
<tr><td colspan="2">Название проекта<br>Теоретическое и техническое обеспечение корпуса текстов на языке пулар.</td></tr>
<tr><td colspan="2">Научный руководитель проекта (ФИО полностью, уч. ст.)<br>Коваль Антонина Ивановна, к.ф.н.</td></tr>
<tr><td>E-mail, телефон, факс (с кодом города)<br>почтовый адрес руководителя проекта</td><td>aekibrik@gmail.com</td></tr>
<tr><td rowspan="3">Полное и краткое название организации – адресата финансирования</td><td>ФИО (полностью) руководителя организации – адресата финансирования<br>Алпатов Владимир Михайлович</td></tr>
<tr><td>ФИО (полностью) главного бухгалтера организации – адресата финансирования<br>Клезович Ирина Ивановна</td></tr>
<tr><td>Телефон, факс (с кодом города), E-mail организации – адресата финансирования<br>Москва, 125009, Б. Кисловский пер. д.1 стр. 1<br>E-mail: iling@iling-ran.ru<br>Тел. дирекции: (495) 690-35-85, Факс: (495) 690-05-28<br>Бухгалтерия: e-mail: 6917875@mail.ru<br>Тел.: (495) 291-78-75, 291-00-63</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>Косогорова М.А.,<br>e-mail: maria.kosogorova@gmail.com</td></tr>
<tr><td>Косогоров В.Н., e-mail: vadimk@rbcmail.ru</td></tr>
<tr><td></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

Направление 4. Создание и развитие корпусных ресурсов по языкам мира.

2. Название проекта

Теоретическое и техническое обеспечение корпуса текстов на языке пулар.

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

Коваль Антонина Ивановна, к.ф.н., ведущий научный сотрудник

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

Косогорова Мария Александровна, ИЯз РАН, МНС.

Косогоров Вадим Николаевич, ООО "Аплана.ЦР", инженер-аналитик.

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

По состоянию на начало 2012 год был готов пилотный проект корпуса аннотированных текстов на языке пулар. Результаты этого проекта, хоть и значительные, продемонстрировали ряд недостатков как в теоретическом обеспечении аннотирования текстов пулар, так и недостатки программного ресурса для глоссирования LightParser. В течение 2012 года основная работа по гранту проходила в двух направлениях, позволяющих скорректировать эти недостатки.

**I. Теоретическое обеспечение.**

Во-первых, принципиальной доработки потребовала система разметки конструкций при изменённом фокусе контраста. Было решено не указывать наличие этого фокуса, равно как и вида глагола, использованного в таких конструкциях. Причиной такого решения является невозможность в 42% случаев с достоверностью определить изменение фокуса контраста без привлечения уникальных специалистов по языку или носителей, что недопустимо для поточной обработки текстов.

Во-вторых, потребовали унификации положения о морфочленении словообразовательных флексий в случае наличия фонетических процессов на морфемном шве. Пилотная версия корпуса отличалась разнообразием подходов при решении этой задачи, что, опять же, недопустимо при введении автоматический системы глоссирования.

В-третьих, был изменён способ аннотации заимствований: предыдущий не был проработан в достаточной мере и не вполне отвечал требованию экономичности.

Также требовал доработки словарь, используемый для работы программы. В частности, была проведена проверка по словарям диалектов с заменой перевода на более общий, разделением, по мере необходимости, переводов на два значения, и также унификацией вида лексемы-перевода.

**II – Внедрение доработанных теоретических принципов**

По мере теоретической доработки принципов глоссирования встала необходимость применить их на практике. Для этого пилотный корпус был переработан с учётом перечисленных изменений, заново снята омонимия, проведён анализ оставшихся недостатков, которые были вынесены в отдельный список на возможную доработку.

**III – Техническая работа**

Второе из основных направлений работы – это модификация программы-парсера а) с учётом изменений, внесённых в теоретическую часть и б) с учётом изменений, необходимых для автоматизации корпуса и разделения ручной работы с корпусом на общую и экспертную.

В рамках первой части были изменения, в основном, были внесены в словарь. Программный код остался фактически неизменённым. Вторая же часть потребовала существенного изменения как кода (см. koval_1) , так и словаря. Во-первых, в эксплуатацию полностью введена система автоматического выбора значения глоссы по части речи. Это – первая ступень планируемой работы по автоматическому снятию омонимии. Для внедрения этой системы каждой единице в словаре (как морфемам, так и независимым лексемам) было присвоено частеречное значение (одно или несколько, см. koval_2), затем подпрограмма была проверена при повторном глоссировании пилотного корпуса (см. koval_3).

Была разработана сложная система поиска по словарям и текстам. На начало 2012 года парсер предоставлял возможность поиска по лексеме пулар в словаре, а также стандартный текстовый поиск. Система поиска позволяет задать область поиска, а также поиск по неполной словоформе, что облегчает задачу работы с большими объёмами текстов.

Также были исправлены некоторые ошибки технического характера, например, возникновение пробела при разбивке ячеек или возникновение лишнего дефиса при дополнительном членении (проводимом уже после подгрузки текста в парсер).

На ноябрь 2012 года также находятся в разработке следующие системы программы-парсера:

- Система двойного маркирования (аналитические формы, рамочные конструкции). Её необходимость обусловлена снятием с эксперта задачи по выделению таких конструкций в тексте. Двойное маркирование появляется при фиксации программой сочетания необходимых частицы и аффикса на заданном расстоянии друг от друга. Тестовый вариант системы проходит проверку. Планируемый срок ввода в эксплуатацию – апрель 2013.
- Программа статистического снятия омонимии. Это программа, после её создания и обучения, позволит принимать автоматические решения по снятию омонимии, основываясь на статистических данных. Для этого корпус должен достичь размеров порядка 25000 словоформ. В первую очередь статистически можно будет снимать местоименную омонимию, затем – глагольную словообразовательную. Планируемый срок готовности программы и начала её обучения – ноябрь 2013.
- Система разметки циркумфикса именного класса. Прототип системы проходит тестирование, однако ему необходимы некоторые дополнения. Эта система позволит соотносить ступень начального согласного корня с аффиксом класса как в существительных, так и в зависимых атрибутах, изменяющихся по классам. Планируемый срок готовности системы – март 2012.

**IV – Программа поиска.**

После долгих дискуссий было решено отказаться от использования для проекта программы searchToo, и, следовательно и преобразования текстов корпуса в формат xml. Было решено разработать в рамках проекта независимую программу для поиска информации в корпусе глоссированных текстов. На состояние к ноябрю 2012 года зафиксированы технические требования к этому проекту, подробности его создания (и возможности их реализации на выбранной платформе ASP .NET) обсуждаются.

Таким образом, результаты работы по проекту в 2012 году относятся к разряду вклада в будущее: теоретические и технические вопросы, решённые в ходе работы, имеют небольшое физическое воплощение, но их результаты позволят в ходе дальнейшей работы избежать задержек и проблем.

12. Важнейшие научные результаты работы по проекту (ок. 0,5 стр. для публикации на сайте Программы)

*В ходе работы были внесены изменения в теоретическое обоснование корпуса, необходимые для упрощения задачи автоматического глоссирования и уменьшения экспертного вмешательства. Изменены принципы разметки прифокусных конструкций, заимствований, словарь унифицирован и выверен с помощью письменных источников.*

*Программа LightParser модифицирована для дальнейшего введения автоматического снятия омонимии. Разработана и введена система автоматического выбора значения глоссы по части речи. Словарь изменён и расширен с учётом системы частеречного сочетания. Разработана система поиска по тексту и словарям, позволяющая задать область поиска, а также поиск по неполной словоформе. Исправлен ряд технических недочётов программы.*

*Создано техническое задание и пилотная версия для системы статистического снятия омонимии для ряда местоименных групп. Создана тестовая версия системы разметки классного циркумфикса. Тестируется и дополняется пилотная система двойного маркирования синтаксических конструкций. Разработано техническое задание для программы поиска в аннотированном корпусе.*

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

*Внесены изменения в теоретическое обоснование корпуса. Изменены принципы разметки прифокусных конструкций, заимствований, словарь унифицирован и выверен с помощью письменных источников. Программа LightParser модифицирована для дальнейшего введения автоматического снятия омонимии. Словарь изменён и расширен с учётом подпрограммы по частеречному конкордансу. Разработана система поиска по тексту и словарям.*

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 планируется:

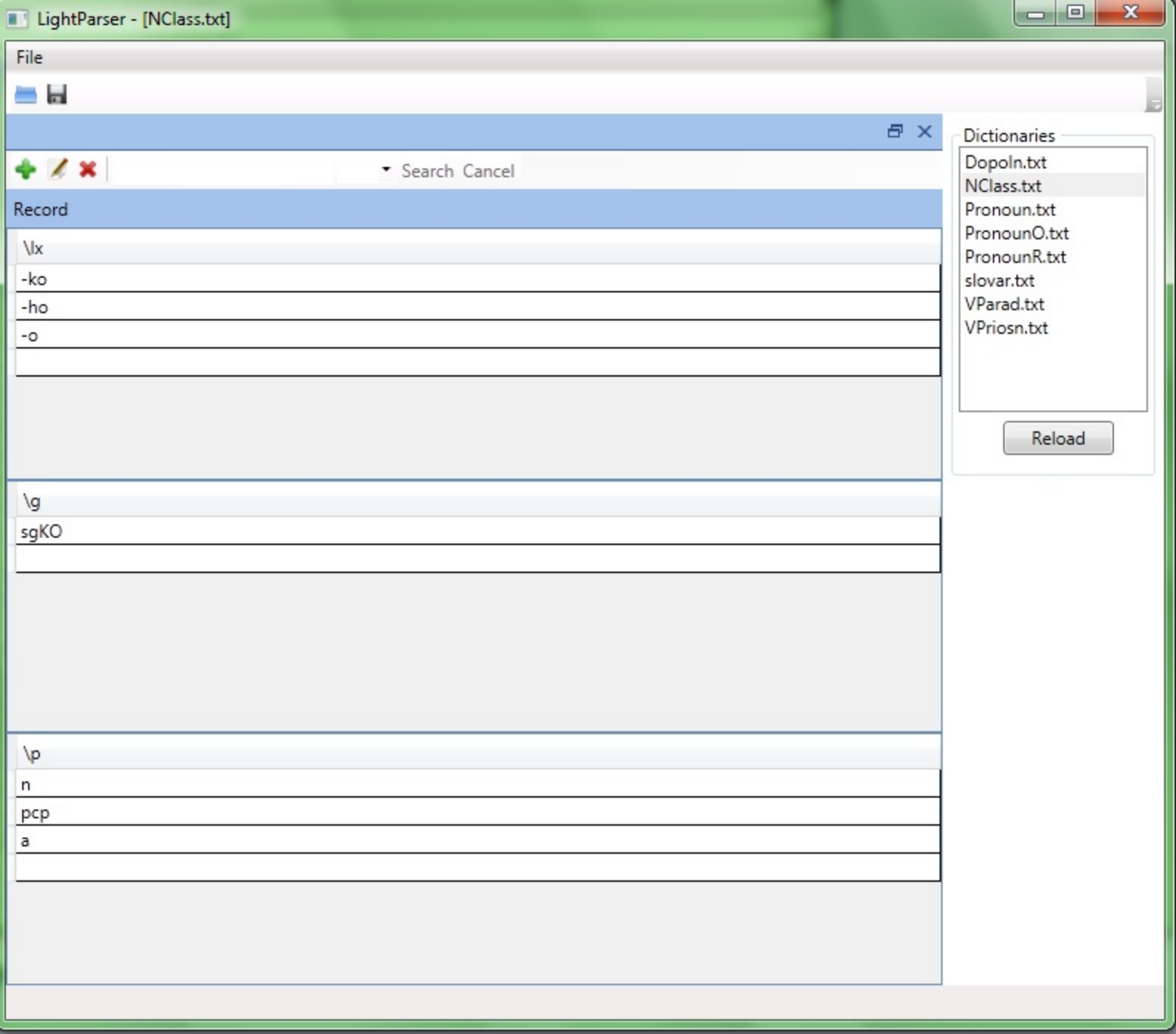
1) Создание программы поиска по тексту корпуса в соответствии с принятыми договорённостями, её тестирование и размещение в открытом доступе.
2) Дополнение программы LightParser системами двойного маркирования, оформления циркумфикса, а также возможностью комментирования.
3) Наполнение корпуса текстами объёмом не меньше 10 000 словоформ и размещение его в открытом доступе.
4) Создание программы подготовки текста к глоссированию и автоматическому наполнению словаря.
5) Разработка уровней представления текстов в глоссировании.

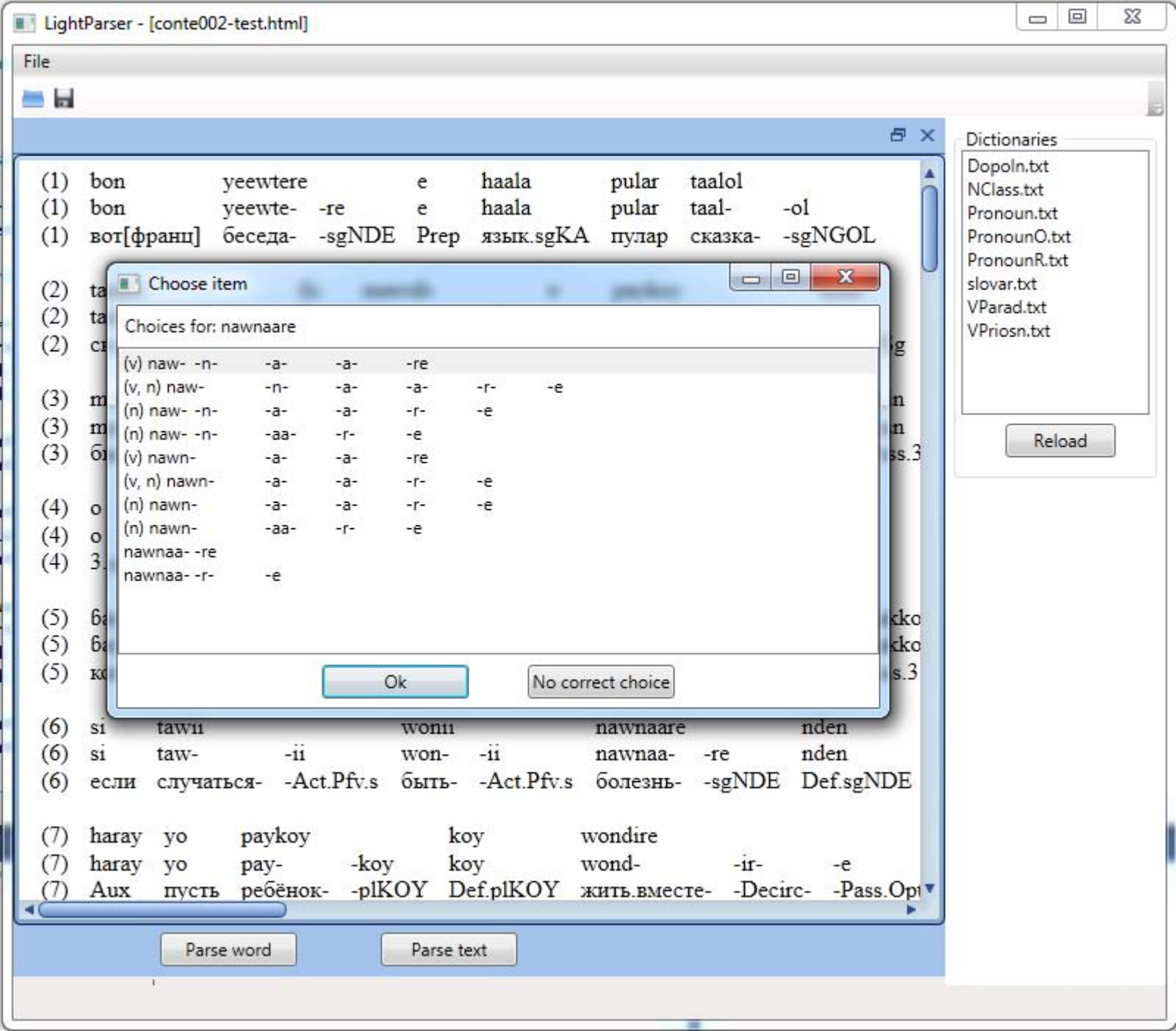
Подпись руководителя проекта А.И. Коваль

**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 |
|---|---|---|---|---|---|
| | Теоретическое и техническое обеспечение корпуса текстов на языке пулар. | ИЯз РАН | Коваль А.И. | | 1) Модификация программы LightParser<br>2) Внедрение программы поиска<br>3) Пополнение корпуса текстов не менее, чем на 10 000 словоформ с размещением его в открытом доступе. |

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;

namespace LightParser
{
    public sealed class Parser
    {
        Singleton

        public IList<Division> DivideWord(string word, List<Record> recordDatabase)
        {
            List<Division> divisions = new List<Division>();
            Division currentDivision = new Division();

            foreach (Record record in recordDatabase)
            {
                if (word == record.Lexeme &&
                    (record.Position == RecordPosition.Any ||
                    record.Position == RecordPosition.BeginningOrWhole ||
                    record.Position == RecordPosition.WholeOnly))
                {
                    currentDivision.Lexemes.Add(word);
                    currentDivision.Records.Add(record);

                    currentDivision.Completed = true;
                    divisions.Add(currentDivision);
                    currentDivision = new Division();
                }
                else if (word.StartsWith(record.Lexeme) &&
                    (record.Position == RecordPosition.Any ||
                    record.Position == RecordPosition.BeginningOrWhole ||
                    record.Position == RecordPosition.BeginningOnly))
                {
                    currentDivision.Lexemes.Add(record.Lexeme);
                    currentDivision.Records.Add(record);

                    // possible parts of speech are at first equal to parts of word's root, which is always the first lexeme to be found
                    currentDivision.PossibleParts.AddRange(record.Parts);

                    List<Division> newDivisions = DivideWordPart(word.Substring(record.Lexeme.Length), currentDivision, recordDatabase);

                    foreach (Division newDivision in newDivisions)
                        if (newDivision.Completed)
                            divisions.Add(newDivision);
                    currentDivision = new Division();
                }
            }

            return divisions;
```


LightParser - [NClass.txt]
File
Search
Cancel
Record
\lx
-ko
-ho
-o
\g
sgKO
\p
n
pcp
a
Dictionaries
Dopoln.txt
NClass.txt
Pronoun.txt
PronounO.txt
PronounR.txt
slovar.txt
VParad.txt
VPriosn.txt
Reload


LightParser - [conte002-test.html]
File
(1) bon yeewtere e haala pular taalol
(1) bon yeewte- -re e haala pular taal- -ol
(1) вот[франц] беседа- -sgNDE Prep язык.sgKA пулар сказка- -sgNGOL
(6) si tawii wonii nawnaare nden
(6) si taw- -ii won- -ii nawnaa- -re nden
(6) если случаться- -Act.Pfv.s быть- -Act.Pfv.s болезнь- -sgNDE Def.sgNDE
(7) haray yo paykoy koy wondire
(7) haray yo pay- -koy koy wond- -ir- -e
(7) Aux пусть ребёнок- -plKOY Def.plKOY жить.вместе- -Decirc- -Pass.Opt
Choose item
Choices for: nawnaare
(v) naw- -n- -a- -a- -re
(v, n) naw- -n- -a- -a- -r- -e
(n) naw- -n- -a- -a- -r- -e
(n) naw- -n- -aa- -r- -e
(v) nawn- -a- -a- -re
(v, n) nawn- -a- -a- -r- -e
(n) nawn- -a- -a- -r- -e
(n) nawn- -aa- -r- -e
nawnaa- -re
nawnaa- -r- -e
Ok
No correct choice
Dictionaries
DopolN.txt
NClass.txt
Pronoun.txt
PronounO.txt
PronounR.txt
slovar.txt
VParad.txt
VPriosn.txt
Reload
Parse word
Parse text