<table>
<tr><td colspan="2">Номер и название направления Программы Корпусная лингвистика. Создание и развитие корпусных ресурсов по языкам народов России</td></tr>
<tr><td colspan="2">Название проекта “Создание корпуса текстов республиканских газет на башкирском языке”</td></tr>
<tr><td colspan="2">Научный руководитель проекта
Хисамитдинова Фирдаус Гильмитдиновна, д.ф.н.</td></tr>
<tr><td>E-mail, факс
почтовый адрес руководителя проекта</td><td>hisamitdinova@list.ru</td></tr>
<tr><td rowspan="3">Полное и краткое название организации — адресат финансирования
Федеральное государственное бюджетное учреждение науки Институт истории, языка и литературы Уфимского научного центра Российской академии наук
ИИЯЛ УНЦ РАН</td><td>ФИО руководителя организации — адресата финансирования
Хисамитдинова Фирдаус Гильмитдиновна</td></tr>
<tr><td>ФИО главного бухгалтера организации — адресата финансирования
Иркина Регина Венеровна</td></tr>
<tr><td>Телефон, факс, E-mail организации
(347)2356077, rihll@anrb.ru</td></tr>
<tr><td>Год начала – год окончания пректа</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="2">Исполнители</td><td>Сиразитдинов Зиннур Амирович, к.ф.н.</td></tr>
<tr><td>Бускунбаева Лилия Айсовна, к.ф.н.</td></tr>
<tr><td>Дата сдачи отчета 19.11.2012</td><td>Подпись руководителя поекта</td></tr>
</table>

Форма 1

УТВЕРЖДАЕМ
Координатор Программы акад. РАН Вяч.Вс. Иванов

Координатор Программы чл.-корр. РАН В.А.Плунгян

“_____” _____ 2012

1. Название направления Корпусная лингвистика. Создание и развитие корпусных ресурсов по языкам народов России

2. Название проекта Создание корпуса текстов республиканских газет на башкирском языке

3. Руководитель проекта (ФИО, ученая степень, должность, место работы)
Хисамитдинова Фирдаус Гильмитдиновна, доктор филологических наук, директор Института истории, языка и литературы УНЦ РАН
4. Основные участники проекта
Сиразитдинов Зиннур Амирович, кандидат филологических наук, старший нучный сотрудник отдела языкознания Института истории, языка и литературы УНЦ РАН
Бускунбаева Лилия Айсовна, кандидат филологических наук, научгый сотрудник отдела языкознания Института истории, языка и литературы УНЦ РАН
5. Содержание фактически проделанной работы, полученные результаты, их новизна, научная и практическая значимость, актуальность, соответствие мировому научному уровню.

I. Определены системы разметок текстов публицистики:

а) Экстралингистические разметки:
дата издания *(для газетных текстов – год, месяц, число, для журнальных – год, месяц, номер выпуска)*;
автор;
название источника;
жанр текста:

- интервью, беседа;
- статья, очерк, репортаж, обозрение;
- рецензия;
- обзор печати;
- советы;
- письма;
- поздравления;
- художественно-публицистические жанры: (эссе, фельетон, рассказ, стихи, эпиграммы);

тематика текста:

- политическая и социальная жизнь (политика, право, философия);
- экономика (бизнес, финансы, коммерция, производство, сельское хозяйство, строительство);
- искусство и культура;
- наука и техника;

- образование;
- природа, путешествие;
- здравоохранение, медицина;
- частная жизнь.

объем текста (число предложений, словоформ);
тип носителя (газета, журнал).

Данный тип разметки дает возможность пользователю, ограничив область поиска и отметив в совокупности или отдельно интересующие его параметры, за короткий промежуток времени получить нужную информацию.

б) Морфологические разметки:
исходная форма слова (лемма),
признак части речи,
признаки грамматических категорий.

Соответственно грамматике башкирского языка морфологическая разметка имеет вид:

N (noun) – существительное: *әсәй, дәрт*

V (verb) – глагол: *уйнау, ярҙамлашыу*

NUM (numeral) – числительное: *биш, йөҙәр*

ADV (adverb) – наречие: *тиҙ, алдан*

ADJ (adjective) – имя прилагательное: *матур, ҡур*

PRON (pronoun) – местоимение: *мин, ниндәй*
POST (postposition) – послелог: *менән, өсөн*
CONJ (conjunction) – союз: *әммә, йәиһә*
PART (particle) – частица: *ғына, әле*
INTJ (interjection) – междометие: *ай, тфу*
PARENTH (parenthesis) – вводное слово: *әлбиттә, ахыры*
PRAEDIC (predicative)– предикатив: *түгел, мөмкин, кәрәк*

Имена существительные

Категория числа

sg (singular) – единств. число: *китап, ҡәләм*

pl (plural) – множ. число: *китаптар, ҡәләмдәр*

Категория падежа

nom (nominative) – номинатив, основной падеж (төп): *ҡағыҙ*

gen (genitive) – генитив, родительный (эйәлек): *ҡағыҙҙың*

dat (dative) – датив, дательный (төбәү): *ҡағыҙға*

acc (accusative) – аккузатив, винительный падеж (төшөм): *ҡағыҙҙы*

abl (ablative) аблатив, исходный (сығана): *ҡағыҙҙан*

loc (locative) – локатив, местный падеж (урын-ваҡыт): *ҡағыҙҙа*

Категория принадлежности

poss1, sg – 1 л. ед.ч.: *уҡыусым*

poss2, sg – 2 л. ед.ч.: *уҡыусың*

poss3, sg – 3 л. ед.ч.: *уҡыусыһы*

poss1, pl – 1 л. мн.ч.: *уҡыусыбыҙ*

poss2, pl – 2 л. мн.ч.: *уҡыусығыҙ*

poss3, pl – 3 л. мн.ч.: *уҡыусылары*

Категория одушевленности/неодушевленности

anim – одушевленность

inan – неодушевленность

Категория сказуемости

pred, 1p, sg – 1 л. ед.ч.: *уҡыусымын*

pred, 2p, sg – 2 л. ед.ч.: *уҡыусыңың*

pred, 1p, pl – 1 л. мн.ч.: *уҡыусыбыҙ*

pred, 1p, pl – 2 л. мн.ч.: *уҡыусыңығыҙ*

Глагол

Категория наклонения

ind (indicative) – изъявительное наклонение: *алдыҡ*

imp (imperative) – повелительное наклонение: *алығыҙ*

int (intentional) наклонение намерения: *алмаҡсымын*

cond (conditional) – условное наклонение: *алһаҡ*

sbjv (subjunctive) – сослагательное наклонение: *алыр ине, алған булыр ине, ала ине, аласаҡ ине*

opt (optative mood) – желательное наклонение: *барғы килә*

Неспрягаемые формы глагола

inf (nfinitive) – инфинитив: *уҡырға*

ptcp (participle) – причастие: *әйткән һүҙ, китеүсе, һөйләшәһе һүҙ*

ger (gerund) – деепричастие: *ауылға барып, оҡата барыу, хәбәр ишеткәс, туйғансы ашаны, әйткәнсә эшлә, өйгә ҡайтышлай*

имя действия: *инеү, барыу менән*

Категория времени

pst def (Past definite tense) – Прошедшее определенное время: *килдем*

pst indf (Past indefinite tense) – Прошедшее неопределенное время: *килгәнмен*

pqpf def – (Plusquamperfect definite tense) – Предпрошедшее определенное время: *килгәйнем*

prs (Present tense) – настоящее время: *киләм*

fut def (Future definite tense) – будущее определенное время: *киләсәкмен*

fut indf (Future indefinite tense) – будущее неопределенное время: *килермен*

Категория отрицания

neg (negation, negative) – отрицание: *килмә*

Местоимение

n-pron – местоимение-существительное: *был, кем*

adv-pron – местоимение-наречие: *шунлыҡтан, ҡайҡа*

adj-pron – местоимение-прилагательное: *ошондай, бындай, беҙҙең*

Прилагательное

Степени сравнения

sup (superlative degree) – *ҡып-ҡыҙыл, ап-аҡ*

com (comparative degree) – сравнительная степень: *ҡурыраҡ, бей*

slac (slackening degree) – степень ослабления: *аҡһыл, алһыу*

Морфологическая разметка в национальном корпусе башкирского языка включает в себя исходную форму слова (лемму), признак части речи, признаки грамматических категорий.

Например, глагол *Ҡара-ма-ған-һың* ‘ты не смотрел’ содержит в себе признаки следующих грамматических категорий: изъявительное наклонение, основной залог, отрицательная форма, прошедшее неопределенное время, 2 лицо, единственное число.

В национальном корпусе данный пример выглядеть таким образом:

лемма: Ҡарау

часть речи**:** V

Семантическая разметка**.** Данная разметка в корпусе башкирского языка предназначена для поиска по лексико-семантическим признакам.

На начальном этапе работы над национальным корпусом башкирского языка планируется внедрение следующих лексико-семантических разрядов и разметок:

Имена существительные (N)

Разряды

r:concr — предметные имена (*йөрәк, ултырғыс, айыу*)

r:abstr — непредметные имена (*Ҡараңғылыҡ, мәнфәғәт, дәрт*)

r:propn — имена собственные (*Сибай, Ҡытырна, Шүлгәнташ)*

Имена собственные

Лексико-семантические пометы

Таксономия:

t:hum | t:hum:supernat — лица (Вәлиди, *Людмила, Черномор*)

t:persn — имена (*Айнур, Зәбилә*)

t:patrn — отчества (*Ғәлиевна*)

t:famn — фамилии (*Баһаутдинов*)

t:topon — топонимы (*Өфө, Ҡырмыҫҡалы, Мәскәү, Ағиҙел*)

Имена прилагательные (A)

Разряды

r:qual — качественные (*матур, ҡыҙыл, һыуыҡ, тәмле*)

r:rel — относительные (*сәскәле, яҙғы, емерек, ташландыҡ*)

Имена числительные (NUM, A-NUM)

Разряды

r:card — количественные (*бер, мең, туғыҙ*)

r:ord — порядковые (*унынсы, йөҙөнсө, һикһәненсе*)

r:distr – разделительные (*берәр, алтышар, етешәр*)

r:coll – собирательные (*берәү, бишәү, йөҙәү*)

r:appr – приблизительные (*бишләп, йөҙләгән*)

r:msr (MEASUREMENT) – меры (*бишле, етеле*)

Местоимения, в том числе:

S-PRO — местоимения-существительные (*мин, кем*)

A-PRO — местоимения-прилагательные (*мине, ниндәй, кемдең*)

ADV-PRO — местоимения-наречия (*ҡайҙа, нисек*)

Разряды

r:pers — личные (*беҙ, ул*)

r:ref — возвратные (*себя*)

r:poss — притяжательные (*уныҡы, быныҡы, минеке*)

r:rel — вопросительные (*нимә, күпме, ҡайһы*)

r:dem — указательные (*бынау, теге*)

r:indet — неопределенные (*ҡасандыр, әллә ниндәй, әллә нисек*)

r:neg — отрицательные (*һис кем, бер кем дә*)

r:spec — кванторные (определительные) (*һәр, һәммә, һәр кем*)

Наречия (ADV)

Лексико-семантические пометы

t:place — место (*алда, бында*)

t:time — время (*иртә, элек, кисә, бая, башта*)

t:manner – образ действия (*йәйәү, аҡсалата*)

уподобление (*байрамса, башҡаса, йәштәрсә*)

t:degree – мера и степень (*байтаҡ, бөтөнләй, икеләтә*)

t:cause – причина и цель (*аңғарматҡан, күрәләтә, юрамал*)

Глаголы (V)

t:move — движение (*барыу* ‘идти’, *һикереү* ‘прыгать’);

t:speech — речь (*әйтеү* ‘сказать’, *өндәшеү* ‘обращаться (к кому)’);

t:psych — психическое состояние (*шатланыу* ‘радоваться’, *аптырау* ‘удивляться’);

t:ment — мышление (*аңлау* ‘понять’, *фекерләү* ‘мыслить’);

t:perc — чувственное восприятие (*ҡарау* ‘смотреть’, *һиҙеү* ‘чувствовать’);

t:physiol — физиологическая сфера (*һауығыу* ‘выздоравливать’, *йоҡлау* ‘спать’);

t:action — действие (*яҙыу* ‘писать’, *төҙөү* ‘строить’);

t:weather — природное явление (*буранлау* ‘мести (о метели)’, *йәшнәү* ‘сверкать (о молнии)’.

Более развернутая система лексико-семантической информации, которая будет включать в себя расширенный тематический класс и словообразовательные характеристики лексемы будет разработана в ходе дальнейших работ.

в) Семантические разметки:

Включает следующие лексико-семантические разряды:

Имена существительные (N)

Разряды

concr — предметные имена (*йөрәк, ултырғыс, айыу*)

abstr — непредметные имена (*ҡараңғылыҡ, мәнфәғәт, дәрт*)

propn — имена собственные (*Сибай, Аҡтырнаҡ, Шүлгәнташ)*

Имена собственные

Лексико-семантические пометы

persn — имена (*Айнур, Зәбилә*)

patrn — отчества (*Ғәлиевна, Хәбирович*)

famn — фамилии (*Баһаутдинов, Ноғаманов*)

topon — топонимы (*Өфө, Ҡырмыҫҡалы, Мәскәү, Аҡъегет*)

Имена прилагательные (**A**)

Разряды

qual — качественные (*матур, ҡыҙыл, һыуыҡ, тәмле*)

rel — относительные (*сәскәле, яҫы, емерек, ташланды*ҡ)

Имена числительные (**NUM**)

Разряды

card — количественные (*бер, мең, ту□һан*)
ord — порядковые (*унынсы, йө□өнсө, һикһәненсе*)
distr – разделительные (*берәр, алтышар, етешәр*)
coll – собирательные (*берәү, бишәү, йө□әү*)
appr – приблизительные (*бишләп, йө□ләгән*)
msr– меры (*бишле, етеле*)

Местоимения (PRON)

Разряды

pers — личные (*бе□, ул*)
poss — притяжательные (*уны□ы, быны□ы, минеке*)
rel — вопросительные (*нимә, күпме, □айһы*)
dem — указательные (*бынау, теге*)
indet — неопределенные (*□асандыр, әллә ниндәй, әллә нисек*)
neg — отрицательные (*һис кем, бер кем дә*)
spec — кванторные (определительные) (*һәр, һәммә, һәр кем*)

Глаголы (**V**)

Лексико-семантические пометы

move — движение (*□у□ғал, тәгәрә, бар, һикер*)
speech — речь (*әйт, өндәш, са□ыр, □абатла*)
psych — психическое состояние (*оял, шатлан, аптыра*)
ment — мышление (*аңла, фекерлә, төшөн)*
perc — чувственное восприятие (*□ара, һи□, ишет*)
physiol — физиологическая сфера (*һауы□, йо□ла, ят*)
action — действие (*я□, тө□ө, □у□ғат*)
weather — природное явление (*буранла, йәшенлә, яу*)

Наречия (**ADV**)

Лексико-семантические пометы

place — место (*алда, бында*)
time — время (*иртә, элек, кисә, бая, башта*)
manner – образ действия (*йәйәү, а□ҡалата*)
likening – уподобление (*байрамса, баш□аса, йәштәрсә*)
degree – мера и степень (*байта□, бөтөнләй, икеләтә*)
cause – причина и цель (*аңғарма□тан, күрәләтә, юрамал*)

Частицы (**PART**)

Разряды

lim — ограничительные (*ғына/генә, □ына/кенә*)
intens – усилительные (*у□, үк; та/тә, да/дә, □а/□ә, ла/лә*)
conf – подтверждения (*дабаһа/□абаһа/лабаһа/табаһа*)
indef – неопределенности (*әле*)

Модальные слова (**MOD**)

Разряды

affir/neg — утверждение/отрицание (*эйе, ю□, түгел, әлбиттә*)
necess – необходимость (*кәрәк, тейеш)*
hypoth– вероятность, гипотетические (*ахыры, бәлки, күрәһең*)
possib – возможность (*мөмкин, ярай*)
induc – побуждение (*зинһар, рәхим итеге□, □уй*)

Подражательные слова (IMT)

echoic – звукоподражательный (*шалтыр, кетер, былт*)
image – образоподражательный (*был□, йым-йым*)

Союзы (CONJ)
Разряды

coord – сочинительный (*һәм, ләкин*)
copul – соединительный (*һәм, да/дә, тағы, йәнә, шуға күрә)*
advers – противительный (*ләкин, юғиһә, фә□әт*)
disjun – разделительный (*йә, хатта, берсә*)
subord – подчинительный (*әгәр, сөнки, тимәк*)
expound – изъяснительный (*тимәк, йәғни*)
causal – причинные (*сөнки, шуға күрә*)
cond – условный (*әгәр, әгәр □ә*)
comp – сравнительный (*әйтерһең, әйтерһең дә*)

Послелоги (POST)
Разряды

nom – управляющий основным падежом (*менән, өсөн, кеүек*)
dat – управляющий дательным падежом (*табан, □арай, са□лы)*
abl – управляющий исходным падежом (*һуң, баш□а, бирле)*

II. Разработан алгоритм и программа программа автоматического анализа башкирской словоформы.

Для агглютинативного башкирского языка предложен подход в многомерном (тензорном) представлении. Любая основа башкирского языка представляется в виде элемента многомерного объекта:
$^{pt}_{s}a^{f}_{i}$ – i-я основа, где p, t – определяют фонетическое строение слова.

f – определяет тип изменений в конечном звуке основы при присоединении аффиксов.

s – определяет принадлежность слова (s=1 — исконно башкирские и ранние заимствованные слова, s=2 — поздние заимствования из русского языка, имеющие значительные отличия при словоизменении (изменения в корне и в типах аффиксов словоизменения, например: *лагерь: лагер□ан, лагерға; волость: волос□а, волостан, волосы*); s=3 — заимствования из арабского и персидского языков, имеющие отличия при словоизменении.

Вводится оператор сцепления текстовых переменных ⊗. Данный оператор сцепляет только те элементы псевдотензора, которые имеют одинаковые значения индексов p, t. Текстовыми переменными для данного оператора являются корень и аффикс, либо корень + аффикс + аффикс и т.д.

Тогда алгоритм анализа башкирской словоформы состоит из следующих процедур:

1. Задается линейный список возможных в языке моделей словоизменений ($A^{jklmnvuxyz}$): 611 для именных и 665 для глагольных форм.

2. Для удобства индексации и последующего анализа словоформы вводится двухмерный массив. Столбцы массива определяют грамматические категории, а подкатегории определяются строками данного массива (i, j, k и т.д.). Фонетическая строение конкретных аффиксов описывается как элемент массива в виде матрицы 4x4 (p и t).

3. Программа посимвольно выделяет слева возможный корень словоупотребления, сравнивая со словарем основ, определяет его фонетическое строение.

4. В процессе вычленения аффиксов от остатка строится дерево возможных сочетаний аффиксов словоизменения для данной основы. При добавлении нового аффикса к узлу дерева проверяется его конечность.

5. Если словоформа полностью разложена, то разложение сравнивается с заданной моделью словоизменения.

6. При наличии нескольких возможных основ, расматриваются все возникающие варианты.

Реализованная программа анализа башкирской словоформы показывает удовлетворительную работу и используется нами в для автоматической морфологической разметки корпуса

III. Подготовлены электронные тексты из газет и журналов, издаваемых на башкирском язке.

Для корпуса башкирской публицистики взяты тексты из следующих источников:

а) Газеты

– «Башкортостан» (с 2002 по 2011 гг.);
– «Киске Өфө» (с 2003 по 2012 гг.);
– «Йәшлек» (с 2003 по 2010 гг.).

б) Журналы:

– «Башкортостан кызы» (с 2002 по 2011 гг.);
– «Ағизел» (с 2003 по 2012 гг.);
– «Шоңкар» (с 2003 по 2010 гг.).

Объем подготовленных текстов для корпуса составялет на текеущий момент 4 миллиона словоупотреблений. Ведется работа по выставлению корпуса в сети Интернет по модели корпуса прозаических текстов http://mfbl.ru/bashkorp/korpus.

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

6.1. количество монографий

6.2. количество сборников статей

6.3. количество статей 3.

7. Список опубликованных монографий и сборников статей, с полным указанием выходных данных, а также их краткие аннотации.

8.Список опубликованных по проекту статей

Сиразитдинов Б.З., Сиразитдинов З.А. О моделировании словоизменительной системы и разработке программ автоматического морфологического анализа башкирского языка//Современное казахское языкознание: актуальны вопросы прикладной лингвистики. Алматы, 2012,с. 103-107.

Сиразитдинов З. А., Бускунбаева Л. А., Ишмухаметова А. Ш., Ибрагимова А. Д., Мигранова Л. Г. Корпус текстов периодической печати на башкирском языке//Актуальные проблемы диалектологии народов Росси: Материалы XII региональной конференции. Уфа, 2012, с. 139-141.

9. Список книг, сданных в печать или поданных на издательские гранты

10. Экспедиции, организованные в рамках проекта

11. Конференции, организованные в рамках проекта.

12. Важнейшие научные результаты работы по проекту. Определена структура и функциональные возможности корпуса башкирских публицистических текстов,разработаны системы лингвистических и экстралингвистических разметок для, создана программа автоматического анализа и разметки текстов, подготовлены сами материалы для корпуса

13. Наиболее значимый научный результат проекта

На базе многомерного представления для агглютинативного башкирского языка разработана система автоматического анализа словоизмененительных форм и программа автоматическокой разметки текстов.

14. Краткий финансовый отчет за 2012 г.

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 форма2

Подпись руководителя

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
	“Создание корпуса текстов республиканских газет на башкирском языке”	Федеральное государственное бюджетное учреждение науки Институт истории, языка и литературы Уфимского научного центра Российской академии наук ИИЯЛ УНЦ РАН	Хисамитдинова Ф.Г. руководитель проекта 5 исполнителей		1. Увеличение объема корпуса на 6 миллионов словупотреблений 2. Увеличение объема базового словаря морфоанализатора на 10 тыс. единиц. 3. Функционирование корпуса в сети Интернет.