ISSN 2079-9160

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ОТДЕЛЕНИЕ ИСТОРИКО-ФИЛОЛОГИЧЕСКИХ НАУК
РОССИЙСКИЙ КОМИТЕТ ТЮРКОЛОГОВ

Российская ТЮРКОЛОГИЯ

Учредители: Институт языкознания РАН
Российский комитет тюркологов при ОИФН РАН

Издатель: от имени РКТ при ОИФН РАН – И.В. Кормушин

№ 2(5)

МОСКВА – КАЗАНЬ 2011

ISSN 2079-9160

COMMITTEE OF RUSSIAN TURCOLOGISTS
RUSSIAN ACADEMY OF SCIENCES
BRANCH OF HISTORY

RUSSIAN TURKOLOGY

Founded by Institute of Linguistics of the Russian Academy of Sciences
The Committee of Russian Turkologists, Russian Academy of Sciences, Branch of History

Publisher: on behalf of the Committee of Russian Turkologists – I.V. Kormushin

№ 2(5)

MOSCOW – KAZAN 2011

Российская тюркология

№ 2(5) 2011

МОРФОЛОГИЧЕСКАЯ РАЗМЕТКА КОРПУСА ХАКАССКОГО ЯЗЫКА

А.В. Шеймович, г. Москва

Резюме. В статье освещается ход работы по созданию корпуса текстов хакасского языка и снабжения его научным аппаратом, а именно – инструментами морфологического анализа. Эта работа ведется в рамках масштабного проекта по созданию корпусных ресурсов языков народов России, в частности корпусов на малых тюркских языках РФ[1], к которым и относится хакасский язык.

Ключевые слова: корпус, корпусная лингвистика, аннотирование корпуса, морфологическое аннотирование, морфологическая разметка, морфологический анализатор.

1.0. О проекте создания хакасского корпуса

Корпусная лингвистика – одно из наиболее перспективных направлений в современной прикладной лингвистике. Оно основано на использовании языковых электронных корпусов и корпусных технологий. Корпус языка – это собрание в электронной форме текстов на данном языке, специально отобранных составителями так, чтобы создать представительную выборку текстов различных жанров и направлений. Языковой материал, организованный в виде корпуса, позволяет максимально быстро получить ответы на вопросы, требующие обработки значительных массивов текстов, ускорить эту обработку и многократно повысить эффективность исследования. Он позволяет решать такие задачи, которые в прошлом лингвисты практически не ставили в силу их трудоемкости или невыполнимости. Это различные виды квантитативных исследований или, например, задачи, связанные с диахроническим исследованием языка. В идеале корпус языка – это собрание не только текстов, но и звучащей речи: это могут быть полевые записи диалектов, аудио- и видеоматериалы (фольклор, записи лекций, теле- и радиопередач). Например, в корпусе может быть представлен не только текст героического сказания, но и его звучащая запись в аутентичном исполнении.

Хакасский корпус, как и большинство корпусов проекта, создается впервые, до сих пор хакасский язык никакого электронного фонда не имел, поэтому сейчас еще нельзя говорить о широком охвате текстов на этом языке. Материалом для корпуса пока служат литературные тексты художественного жанра,

[1] Работа выполнена при поддержке Программы фундаментальных исследований РАН «Корпусная лингвистика. Создание и развитие корпусных ресурсов по языкам народов России» (направление – создание и размещение в интернете корпусов текстов на тюркских языках России, разработка корпусных технологий).

оцифрованные и приведенные к стандартному формату: героическое сказание «Алтын Тайчы», героические эпосы «Ай-Хуучин», «Алтын Арығ», богатырское сказание «Ах Чібек Арығ»; хакасские народные сказки и легенды, пословицы и поговорки; повести, рассказы, пьесы и стихи хакасских писателей и поэтов А. Халларова, В. Кобякова, В. Татарова, В. Майнашева, В.Г. Шулбаевой, Г. Кичеева, М. Баинова и др.; газета «Хакас чірі» («Хакасская земля»).

В распоряжении составителей корпуса есть также оцифрованная версия Большого хакасско-русского словаря на 22 тыс. слов под ред. О.В. Субраковой.

1.1. Что такое морфологическая разметка корпуса

Параллельно с построением корпуса начата работа по созданию морфологического анализатора (или парсера) для хакасского языка. Эта работа также является частью более масштабной задачи. Как известно, эффективным рабочим инструментом корпус может являться только при условии его лингвистического аннотирования (то есть в нем должны быть указаны морфологические, синтаксические, семантические и иные свойства сегментов текста). Мы начинаем свою работу с морфологического аннотирования (разметки). Такая разметка позволит быстро находить в корпусе материал, удовлетворяющий запросу пользователя. Например, можно будет автоматически определить принадлежность словоформы к той или иной лексеме, понять, с каким множеством основ употребляется тот или иной аффикс. Подобная информация может пригодиться для решения широкого круга вопросов, выходящих за рамки грамматики. В перспективе перед нами стоит задача создания универсального морфологического анализатора для тюркских языков. Поэтому создание морфологического анализатора для хакасского можно рассматривать как первый шаг на пути к ее выполнению.

Морфологическая разметка текста состоит в выделении словоформ и в приписывании каждой из них информации об исходной форме слова и о совокупности ее грамматических признаков.

Морфологическая информация, приписываемая словоформе, состоит из следующих групп помет:

1. Основа, которой принадлежит словоформа (указывается «словарный вид» лексемы);
2. Принадлежность лексемы к той или иной части речи;
3. Множество грамматических признаков данной словоформы, или словоизменительные характеристики (напр., падеж существительного, время глагола) [см.: Ляшевская и др. 2005: 114].

Морфологическая разметка содержит информацию о словоизменительных, но не о словообразовательных признаках лексемы. Иначе говоря, программа-парсер будет вычленять в слове лишь те аффиксы, которые имеют грамматическое значение (падеж, лицо, число, время, наклонение, принадлежность и т.п.) и пренебрегать теми, которые выполняют деривационную и семантическую функцию. Например, в слове *палыхчыларыбыстың* ‘наших рыбаков’ словообразовательным является афф. *-чы* (показатель деятеля), он рассматривается как часть основы и не учитывается при морфологическом анализе Парсер проанализирует слово как

палыхчы-лар-ыбыс-тың
рыбак - афф. мн.ч. - афф. принадл. 1л. мн.ч.-афф. род.п.

1.2. Характеристики языка, релевантные для морфологической разметки

Особенности морфологической разметки вытекают из характеристик языка. Типологически хакасский язык принадлежит к языкам агглютинативного типа, т.е. обладает признаками, которые кажутся весьма привлекательными для разработчика морфологического анализатора (особенно в сравнении с языками флективного типа). Вот основные из них:

1) развитая система словоизменительных аффиксов, большинство из которых грамматически однозначно, т.е. одним аффиксом выражается один грамматический признак. Случаи совмещения в одном аффиксе двух значений отсутствуют, однако наблюдается незначительная грамматическая омонимия (см. ниже);

2) единый тип склонения и спряжения – в отличие от флективных языков типа русского и других европейских, где слова, принадлежащие к одной части речи, изменяются весьма различными способами. Например, при создании французского спеллера было выделено более 200 словоизменительных классов глаголов в рамках трех групп, фиксируемых традиционной французской грамматикой;

3) отсутствие значимых чередований в основах, четкая фонетическая обусловленность использования алломорфов.

Иначе говоря, к основе в строгом порядке присоединяются однозначные стандартные аффиксы, границы морфем отчетливы, фонетические изменения на стыках морфем подчиняются строгим правилам. Однако вышеперечисленные «конструктивные достоинства» агглютинативного языка компенсируются количеством упомянутых межморфемных сандхи (фонетических изменений на стыках морфем), о которых речь пойдет чуть ниже, а «развитая система словоизменительных аффиксов» и их грамматическая однозначность приводят к тому, что, например, парадигма имени состоит более чем из 600 мест. Конечно, не все формы употребительны, не все реализуются с одинаковой частотой (частотных – не более 100 для имен существительных), но это – разрешенные для данного языка цепочки морфем, которые автоматический анализатор обязан учитывать.

2.0. Этапы создания морфологического анализатора

2.1. Построение словаря морфем

Для построения морфологического анализатора требуется провести инвентаризацию всех словоизменительных морфем языка. Совокупность аффиксов, выражающих все грамматические категории, существующие в языке, составит инвентарь (словарь) морфем. Список морфем извлекается вручную из грамматики языка. Каждому аффиксу в словаре должно быть приписано свое грамматическое значение (напр., *-лар* – афф. мн. числа, *-да* – афф. местного падежа, *-бын* – афф. 1.л. ед.числа …) и отмечено, с каким типом основ и других аффиксов сочетается данный аффикс.

Корпус словоизменительных аффиксов, выделенных на текущем этапе работы, см. в Приложении (с. 57–61).

2.2. Построение словаря основ

В соответствие словарю морфем ставится словарь основ, в котором будут содержаться полнозначные слова (основы) в начальной (словарной) форме – леммы. Словарь основ автоматическим образом извлекается из Большого хакасско-русского словаря (Новосибирск, «Наука», 2006) с использованием технологий системы управления базами данных StarLing. Импорт текстового файла словаря в StarLing и поэтапное создание на его основе многоуровневой лексико-грамматической базы данных описаны в работе С.А.Крылова «Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии» в разделе «Преобразование текста в стандартно организованную базу данных» [Крылов 2008: 649–650]. С использованием технологий той же системы StarLing планируется конвертировать в базу данных и инвентарь морфем хакасского языка [Крылов 2008, с. 650].

2.3. Операции над словарями основ и морфем

Построение морфологического анализатора сводится к следующим формальным операциям над словарями основ и морфем:

2.3.1. Выделение в «словаре основ» частей речи в соответствии с типами словоизменения

Согласно сложившейся в тюркологии грамматической традиции мы (вслед за Н.А.Баскаковым [Хакасско-русский словарь, 1953]) выделяем в хакасском языке три части речи: имена, глаголы и незнаменательные части речи (частицы, послелоги, союзы и т.п.).

Все части речи (за исключением незнаменательных, являющихся неизменяемыми) характеризуются набором грамматических признаков, каждый из которых выражается определенными формальными показателями (аффиксами). Совокупность этих показателей определяет тип словоизменения и словоизменительный класс данной части речи[2]. Например, для хакасского глагола грамматические признаки – это лицо, число, наклонение, время и т.п.; для имени – число, принадлежность, падеж, лицо. Конечно, это не значит, что в хакасском совершенно отсутствует общепринятое деление имени на более мелкие разряды: существительное, прилагательное, наречие (которое Н.А.Баскаков причисляет к разрядам имени), числительное, местоимение. Однако дифференциация между именными разрядами выражена слабо, особенно между существительными, прилагательными и наречиями: одно и то же слово может трактоваться как существительное, прилагательное или наречие в зависимости от синтаксической функции, выполняемой им в предложении: существительное может выступать в роли определения: *тас тура* ‘каменный дом’; прилагательное может выполнять в предложении любую функцию: *кічіглер ойнапчалар* ‘маленькие (малыши, дети) играют’; многие наречия являются застывшими падежными формами имен (*таңда* ‘завтра’ – местный п. от *таң* ‘заря, рассвет’); числительное, приняв субстантивный словообразовательный аффикс, становится существительным, либо, присоединив атрибутивный аффикс, – прилагательным: *ӱс* ‘три’ – *ӱзӧлең* ‘трое’ – *ӱзінҷі* ‘третий’. Главным же с точки зрения автоматиче-

[2] Как уже было сказано выше, для агглютинативных языков каждой части речи (и соответствующему ей типу словоизменения) в большинстве случаев соответствует лишь один словоизменительный класс.

ского морфоанализа является то, что при таком переходе слов из разряда в разряд они сохраняют единый тип склонения, что дает формальное основание морфологическому парсеру считать их принадлежащими к одному словоизменительному типу и к одной части речи.

2.3.2. Классификация именных и глагольных основ из «словаря основ» в соответствии с их фонетическими характеристиками, которые определяют фонетический облик следующих за основами словоизменительных аффиксов[3].

Основные фонетические закономерности

Важной фонетической закономерностью хакасского языка, отражающейся на нашей работе, является закон сингармонизма, действующий в большинстве агглютинативных языков. Этот закон заключается в уподоблении гласных (а иногда согласных) в рамках одного слова по одному или нескольким фонетическим признакам, таким, как ряд, подъём или огубленность. Сингармонизм по ряду и огубленности характерен для вокализма большинства тюркских языков.

В хакасском языке сингармонизм проявляется в том, что все гласные слова уподоблены по ряду гласному первого слога основы и могут быть либо заднеязычными (*а, ы, у, о*), например: *харах-тар-ыбыс-та* ‘в наших глазах’, либо переднеязычными (*і(и), е, ӱ, ӧ*) *кирек-тен-деңер* ‘по делам’. Гласные основы оказывают влияние и на качество согласных в слове (более переднюю или более заднюю их артикуляцию). Например, в хакасских словах согласные *х, ғ* употребляются только с задними гласными (*хыр* ‘крыша’, *торғы* ‘шёлк’), а согласные *к, г* – только с передними (*кӱн* ‘день’, *тігі* ‘тот’).

В хакасском присутствует также сингармонизм по признаку огубленности, заключающийся в том, что гласные в слове могут быть либо губными (*у, ӱ, о, ӧ*), либо негубными (*а, е, ы, і (и)*). Однако эта закономерность в хакасском проводится непоследовательно и распространяется только на основу слова и то лишь по линии узких гласных, не затрагивая аффиксы **(***пулуң* ‘угол’ – *пулуңы* (вин. п.); *хузурух* ‘хвост’ – *хузурухты* (вин.п.)**)**, а потому губной сингармонизм нерелевантен для выделения типов основ, диктующих выбор фонетических вариантов словоизменительных морфем.

Кроме сингармонизма в языке имеется еще ряд фонетических закономерностей, релевантных для морфологического анализатора. К ним относятся:

1) Выпадение узких гласных (*і, и, ы, у, ӱ*) в позиции перед согласным *н* после присоединения аффикса на гласные *ы, і*, например: *пурун* ‘нос’ – *пурны* < *пуруны* ‘его нос’;

2) Выпадение в интервокальной позиции конечного ***г, ғ*** основы перед гласным словоизменительного аффикса принадлежности, например: *суғ* ‘вода’ – *суу* (< *суғу*) ‘его вода’; *көг* ‘песня’ – *көө* ‘его песня’ (< *көгі*);

3) Выпадение начального *г, ғ,* аффикса после основы на *-г, -ғ, -ң*: *көг* ‘песня’ – *көг-е* ‘песне’ (< *көг-ге*); *суғ* ‘вода’ – *суғ-а* ‘воде’ (< *суғ-ға*); *таң* ‘заря’ – *таңа* ‘заре’ (< *таң-ға*);

4) Ассимиляция согласных, как в основах, так и в аффиксах (после глухих – глухие, после звонких – звонкие):

[3] Необходимо оговориться, что здесь мы рассматриваем только те фонетические закономерности, которые проявляются орфографически, т.к. морфологический анализатор работает только с письменными текстами.

а) прогрессивная ассимиляция (уподобление последующего согласного предыдущим) ярко проявляется в вариативности звуков *л/н/т* в аффиксах мн.ч.:

- после основ на гласные (*а, ы, і, е, о, у, ӧ, ӱ*) и на звонкие согласные *-г, -ғ, -й, -л, -р* употребляется афф. мн.ч. *-лар/-лер* (*тағ-лар, кізі-лер*);
- после хакасских основ на глухие (*-т, -с, -х, -к* и т.д) и заимствованных из русского языка основ на звонкие (*-б, -в, -г, -д, -ж, -з*) употребляется афф. мн.ч. *-тар/-тер* (*ат-тар, хус-тар, завод-тар*)
- после основ на сонорные *-м, -н, -ң* употребляется афф. мн.ч. *-нар/-нер* (*хум-нар, кӱн-нер*);

б) к явлениям регрессивной ассимиляции относится озвончение глухих согласных *к, т, с, п, х* в интервокальной позиции на границе основы и аффикса, например: *ат* ‘имя’ – *ады* ‘его имя’; *сас* ‘волос’ – *сазым* ‘мой волос’; *тап* ‘находить’ – *табыл-* ‘быть найденным; найтись’.

Примеры выделенных типов основ

В соответствии с вышеперечисленными фонетическими закономерностями были выделены несколько типов основ, сочетающихся с различными фонетическими вариантами словоизменительных аффиксов. Вначале основы были разделены по сингармоническому принципу на две группы: заднерядные и переднерядные; затем основы, входящие в каждую из этих групп были классифицированы по типу конечного звука – гласного или согласного, требующего после себя различного звукового оформления всей цепочки словоизменительных аффиксов.

1. Основы, заканчивающиеся на заднерядные гласные и сочетания «заднерядный гласный + разные типы согласных»

1.0. Основы на заднерядный гласный (Vback) *а, ы, о, у* (*тура*);

1.1. Основы на заднерядный гласный + носовой *м, н* (Vback + *м, н*) (*чон*) + чередование в основе, т.н. беглый гласный, напр. *мойын/мойн-*;

1.2. Основы на заднерядный гласный + носовой *ң* (Vback + *ң*) (*таң*);

1.3. Основы на заднерядный гласный + сонант *р, л, й* (Vback + *р, л, й*) (*чар, хол, тахпай*);

1.4. Основы на заднерядный гласный + *ғ* (Vback + *ғ*) (*суғ*) + чередование в основе, напр. *суғ*/*су* интервокальной позиции;

1.5. Основы на заднерядный гласный + глухой *т, п, х, ш, с, ф, ч, ц* [либо оглушающийся *б, в, г, д, ж, з* русских заимствований] (Vback + *т, п, х, ш, с, ф, ч, ц*) (*ат, харах, колхоз*) + чередование в основе – озвончение глухих в интервокальной позиции: *п/б, с/з, х/ғ, т/д* в случае начального гласного аффикса.

2. Основы на переднерядные гласные и сочетания «переднерядный гласный + разные типы согласных»

2.0. Основы на переднерядный гласный (Vfront) *е, и, і, ӧ, ӱ* (*төге*);

2.1. Основы на переднерядный гласный + носовой *м, н* (Vfront + *м, н*) (*кӱн*, *килин*) + чередование в основе, т.н. беглый гласный, напр. *килин/килн-*;

2.2. Основы на переднерядный гласный + носовой *ң* (Vfront + ң) (*төң*);

2.3. Основы на переднерядный гласный + сонант *р, л, й* (Vfront + р, л, й) (*көл*);

2.4. Основы на переднерядный гласный + *г* (Vfront + *г*) (*көг*) + чередование в основе, напр. *көг*/*кө* в интервокальной позиции;

2.5. Основы на переднерядный гласный + глухой *к, т, п, ш, с, ф, ч, ц* [либо оглушающийся *б, в, г, д, ж, з* русских заимствований] (Vfront + *к, т,*

п, ш, с, ф, ч, ц) (*түӱк*) + чередование в основе – озвончение глухих в интервокальной позиции: *п/б, с/з, к/г, т/д* в случае начального гласного аффикса.

Для некоторых из перечисленных случаев потребуется «размножение» словарных основ в словаре с учетом фонетических чередований [напр., интервокального озвончения (пп. 1.5, 2.5) – *пас-/паз-* (*пас* ‘голова’ (им.п., ед.ч.) – *пазым* ‘моя голова’ им.п., афф. принадл. 1л. ед.ч.) или выпадения узкого гласного основы перед *н* (пп. 1.1, 2.1): *пурун-/пурн-* (*пурун* ‘нос’ (им.п., ед.ч.), *пурны* ‘его нос’ им.п., афф. принадл. 3 л. ед.ч.)].

2.3.3. Постановка в соответствие каждому из выделенных типов основ комплекта служебных морфем

Каждому из выделенных фонетических типов основ ставится в соответствие определенный набор словоизменительных аффиксов из словаря морфем со всеми их значениями (омонимичным морфемам д.б. приписаны все возможные значения).

Если на вокализм аффиксов в хакасском языке, исходя из закона сингармонизма, влияет главным образом основа слова, то на их консонантные варианты, согласно правилам прогрессивной ассимиляции, помимо основы оказывают влияние также и предшествующие аффиксы. В хакасском языке, как и в других агглютинативных языках, существует жесткий порядок следования морфем. К корню сначала присоединяются аффиксы словообразования – лексического, а затем грамматического, – затем к основе присоединяются словоизменительные аффиксы в следующем порядке (для имени): аффиксы числа, принадлежности, падежей, лица, например:

тура - ӌах - тар - ыбыс - ха ‘нашим домикам’
дом – афф.– мн.ч.– принадл.–дат. падеж
уменьш. 2л.ед.ч.
(лекс. сл.-обр.)

Поэтому помимо сочетаемости основ и аффиксов должна быть проработана фонетическая сочетаемость аффиксов между собой для каждого возможного их комплекта.

Данные о сочетаемости каждого аффикса с двумя соседними – справа и слева, – как и данные о сочетаемости аффиксов с основами разного фонетического типа вносятся в словарь морфем языка.

Таблицы примеров сочетаемости типов основ и фонетических вариантов аффиксов, а также сочетаемости аффиксов друг с другом приводятся в Приложении.

Введя эти данные в систему StarLing, после процесса обработки текстов корпуса мы сможем получить его морфологическую разметку с грамматической омонимией.

3. Об инструментах создания хакасского корпуса и его морфологической разметки

Отдельно следует сказать о технологиях, применяемых при создании морфологически аннотированного корпуса хакасского языка. Это интегрированная информационная среда StarLing, созданная С.А. Старостиным, во всех отношениях удовлетворяющая задачам корпусной лингвистики, таким как создание и

редактирование лингвистических корпусов, а также снабжение их удобным справочным аппаратом. Этот инструмент уже апробирован при создании систем морфологического анализа для двух языков – русского и английского: на основе «Грамматического словаря русского языка» А.А. Зализняка и англо-русского словаря В.К. Мюллера. В настоящее время система StarLing успешно используется российскими лингвистами при построении корпусов языков различных семей и их морфологической разметке.

Механизмы работы системы StarLing и инструменты, которые она предоставляет составителям корпусов и баз данных, описаны в работах С.А.Крылова [Крылов 2008; 2011].

4. Снятие грамматической омонимии

Для корпуса флективного языка, такого как русский, снятие морфологической омонимии – это довольно большая работа, которая не обходится исключительно автоматическими средствами и требует немалых усилий человека [см.: Ляшевская и др. 2005: 113]. В агглютинативных же языках (в частности в хакасском) грамматическая омонимия сравнительно невелика, однако она есть. Так, например, в именной парадигме существуют омонимичные аффиксы исх. и твор. падежей для основ на гласные и сонорные (*-наң/-нең*); аффиксы напр. падежа для основ на заднеязычные гласные омонимичны аффиксам сказуемости 2 л. мн.ч. для основ того же вида (*-зар/-зер*); омонимичными являются также аффиксы сказуемости и принадлежности 1 и 2 л. мн.ч. для основ на гласные и звонкие согласные (*-м/-ым/-ім*, *-быс/-біс*, *-ң/-ың/-ің, -ңар/-ңер/-ыңар/ -іңер*).

При использовании системы StarLing снятие морфологической омонимии в корпусе становится полуавтоматической процедурой. Если морфологическую омонимию можно разрешить с помощью синтаксического контекста, то «возможен вывод на экран всех записей, удовлетворяющих тому или иному условию, задаваемому в терминах синтаксических фильтров. После такой фильтрации задача снятия морфологической омонимии упрощается (и т.о. убыстряется) в несколько раз, как показывает практика морфологической разметки русскоязычных корпусов» [Крылов 2011: 651].

5. Значение создания корпусов для малых языков России

Создание корпусов малых языков, которых достаточно много на территории Севера, Сибири, Дальнего Востока России, имеет особое значение. Корпус фиксирует уходящие или находящиеся под угрозой исчезновения малые языки и сохраняет их фонд хотя бы для науки и культуры. Корпус – это также способ повысить жизнеспособность и способствовать модернизации относительно крупных языков России, которым вымирание не грозит, но которые, например, недостаточно представлены в интернете (или вообще в нем не представлены). Создание корпуса является хорошим стимулом к расширению сферы употребления языка, его, так сказать, экспансии в интернете. На основе современных кор-

пусных технологий впоследствии можно получить механизмы, без которых немыслима жизнь языка в интернете, такие как поисковые системы, системы автоматического перевода, средства электронной проверки текстов. Хотелось бы отметить, что при полноценной морфологической разметке лингвистического корпуса (в случае его достаточного объема и репрезентативности) в качестве бонуса можно получить электронный спелл-чекер для рассматриваемого языка.

ЛИТЕРАТУРА

Грамматика хакасского языка / Под ред. Н.А.Баскакова. – М., 1975.

Крылов С.А. Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии // Orientalia et classica. Тр. Ин-та восточных культур и античности. Вып. XIX. Аспекты компаративистики. 3/ Ред И.С. Смирнов.– М.: РГГУ, 2008. – С. 649–668.

Крылов С.А. Использование системы StarLing при создании морфологически аннотированного корпуса современного монгольского языка. – *на правах рукописи*.

Ляшевская О.Н., Плунгян В.А., Сичинава Д.В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. – М., 2005. – С. 111–135.

Плунгян В. А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003–2005. – М.: Индрик, 2005. – С. 6–20.

Сиразитдинов З.А. Алгоритмическая грамматика словоизменения башкирского языка / http://mfbl.ru/bashdb/algram/algram.htm

Сичинава Д.В. К проблеме создания корпусов русского языка // Научно-техническая информация, сер. 2, № 11, 2002. – С. 25—31.

Хакасско-русский словарь / Под ред. О.В. Субраковой. – Новосибирск: Наука, 2006.

Хакасско-русский словарь / Под ред. Н.А. Баскакова, с прилож. грамматического очерка хакасского языка Н.А.Баскакова. – М., 1953.

Sheymovich, Aleksandra V. Morphological tagging of the Khakas corpus

Summary. This paper describes the creation of an electronic corpus of the Khakas language and an attempt to develop a morpheme analyzer for Khakas. This research is a part of the project of the Russian Academy of Sciences on the creation of electronic corpora of languages of the Russian Federation, including corpora of Turkic minority languages such as Khakas.

Key words: corpus of a language, corpus linguistics, annotation of a corpus, morphologic annotation, morphologic analyzer

ПРИЛОЖЕНИЯ

Таблица 1

Сочетаемость аффиксов грамматических категорий имени и некоторых фонетических типов основ

Категории и аффиксы		Основы											
		Переднерядные (Vback)						Заднерядные (Vfront)					
		Vback + k	Vback + м, н	Vback + н	Vback + р,л,й	Vback + г	Vback + т,п, х,ш, с,ф,ч,ц	Vfront	Vfront + м, н	Vfront + н	Vfront +р,л,й	Vfront + г	Vfront + к,т, п,ш,с,ф,ч,ц
Афф. мн. числа													
лар		+			+	+							
лер								+			+	+	
нар			+	+									
нер									+	+			
тар							+						
тер													+
Афф. принадл.													
1л. ед.ч.	м	+				+		+				+	
	ым		+	+	+		+						
	ім								+	+	+		+
2л. ед.ч.	ң	+				+			+			+	
	ың		+	+	+		+						
	ің								+	+	+		+
3 л.ед., мн.ч.	ы		+	+	+		+						
	і								+	+	+		+
	зы	+											
	зі							+					
1л.мн.ч.	быс**	+				+							
	біс**							+				+	
	ыбыс		+	+	+		+						
	ібіс								+	+	+		+
2л.мн.ч.	ңар	+				+							
	ңер							+				+	
	ыңар		+	+	+		+						
	іңер								+	+	+		+
Афф. простого склонения													
Осн.	∅												
Род.	ның	+	+	+	+	+							
	нің							+	+	+	+	+	
	тың						+						
	тің												+
Дат.	а	+				+							
	е							+				+	
	ға		+	+	+								
	ге								+	+	+		
	ха						+						
	ке												+
Вин.	ны	+	+	+	+	+							
	ні							+	+	+	+	+	
	ты						+						
	ті												+

Местн.	да	+	+	+	+	+							
	де							+	+	+	+	+	
	та						+						
	те												+
Исходн.	даң	+			+	+		+			+	+	
	дең												
	наң*		+	+									
	нең*								+	+			
	таң						+						
	тең												+
Направ-вит.	зар***	+	+	+	+	+							
	зер***							+	+	+	+	+	
	сар***						+						
	сер***												+
Творит.	наң*	+	+	+	+	+	+						
	нең*							+	+	+	+	+	+
Афф. л. с нулев. п.													
1л. ед.ч.	мын		+	+					+	+			
	мін												
	бын	+			+	+							
	бін							+			+	+	
	пын						+						
	пін												+
2л. ед.ч.	зың	+	+	+	+	+							
	зің							+	+	+	+	+	
	сың						+						
	сің												+
3л. ед.ч.	∅												
1л.мн.ч.	мыс		+	+									
	міс								+	+			
	быс**	+			+	+							
	біс**							+			+	+	
	пыс						+						
	піс												+
2л.мн.ч.	зар***	+	+	+	+	+							
	зер***							+	+	+	+	+	
	сар***						+						
	сер***												+
3л.мн.ч.	∅												

Таблица 2

Сочетаемость аффиксов принадлежности и падежных аффиксов притяжательного склонения

Афф. принадл. / Афф. притяж. склон.		1 л. ед.ч.			2 л. ед.ч.			3 л. ед. и мн. ч.				1л. мн.ч.				2 л. мн.ч.			
		м	ым	ім	ң	ыңң	ің	ы	і	зы	зі	быс	біс	ыбыс	ібіс	ңар	ңер	ңар	ңер
Осн.	∅																		
Род.	ның	+	+		+	+		+		+						+		+	
	нің	+		+	+		+		+		+						+		+
	тың											+		+					
	тің												+		+				
Дат.	а	+	+		+	+													
	е			+	+		+												
	ға															+		+	
	ге																+		+
	ха											+		+					
	ке												+		+				
	на							+		+									
	не								+		+								
Вин.	ны	+	+		+	+										+		+	
	ні	+		+	+		+										+		+
	н							+	+	+	+								
	ты											+		+					
	ті												+		+				
Местн.	да	+	+		+	+													
	де	+		+	+		+									+		+	
	нда							+		+							+		+
	нде								+		+								
	та											+		+					
	те												+		+				
Исходн.	даң															+		+	
	дең																+		+
	наң	+	+		+	+		+		+									
	нең	+		+	+		+		+		+								
	таң											+		+					
	тең												+		+				
Направ.	зар	+	+		+	+										+		+	
	зер	+		+	+		+										+		+
	нзар							+		+									
	нзер								+		+								
	сар											+		+					
	сер												+		+				
Творит.	наң	+	+		+	+		+		+		+		+		+		+	
	нең	+		+	+		+		+		+		+		+		+		+

Таблица 3

Сочетаемость аффиксов лица и падежных аффиксов простого склонения

Афф. лица / Афф. падежа		1 л. ед.ч.					2 л. ед.ч.			3 л. ед. и мн.ч.	1 л. мн.ч.					2 л.мн..ч.		
		мын	мін	бын	бін	пын/ пін	зың	зің	сың/сің	∅	мыс	міс	быс	біс	пыс/ піс	зар	зер	сар /сер
Осн.	∅																	
Род.	ның	+					+				+					+		
	нің		+					+				+					+	
	тың	+					+				+					+		
	тің		+					+				+					+	
Дат.	а			+			+						+			+		
	е				+			+						+			+	
	ға			+			+						+			+		
	ге				+			+						+			+	
	ха			+			+						+			+		
	ке				+			+						+			+	
Вин.	ны			+			+						+			+		
	ні				+			+						+			+	
	ты			+			+						+			+		
	ті				+			+						+			+	
Местн.	да			+			+						+			+		
	де				+			+						+			+	
	та			+			+						+			+		
	те				+			+						+			+	
Исходн.	дан	+					+				+					+		
	ден		+					+				+					+	
	нан*	+					+				+					+		
	нең*		+					+				+					+	
	тан	+					+				+					+		
	тең		+					+				+					+	
Направит.	зар**			+			+						+			+		
	зер**				+			+						+			+	
	сар**			+			+						+			+		
	сер**				+			+						+			+	
Творит.	нан*	+					+				+					+		
	нең*		+					+				+					+	

Таблица 4

Сочетаемость аффиксов лица
и падежных аффиксов притяжательного склонения

Афф. принадл. / Афф. лица		1 л. ед.ч.			2 л. ед.ч.			3 л. ед.ч.				1л. мн.ч.				2 л. мн.ч.			
		м	ым	ім	ң	ың	ің	ы	і	зы	зі	быс	біс	ыбыс	ібіс	ңар	ңер	ыңар	іңер
1 л. ед.ч.	мын	+	+		+	+													
	мін	+		+	+		+												
	бын							+		+						+		+	
	бін								+		+						+		+
	пын											+		+					
	пін												+		+				
2 л. ед.ч.	зың	+	+		+	+		+		+						+		+	
	зің	+		+	+		+		+		+						+		+
	сың											+		+					
	сің												+		+				
3 л..ед. и мн. ч.	∅																		
1л. мн.ч.	мыс	+	+		+	+													
	міс	+		+	+		+												
	быс							+		+						+		+	
	біс								+		+						+		+
	пыс											+		+					
	піс												+		+				
2 л. мн.ч.	зар	+	+		+	+		+		+						+		+	
	зер	+		+	+		+		+		+						+		+
	сар											+		+					
	сер												+		+				

✦✦✦