

## **К морфологической разметке корпуса хакасского языка**

### **Towards a Morphological Annotation of the Khakass Corpus**

А.В. Шеймович  
asheimovich@yandex.ru  
Институт языкознания РАН  
М.Н.С.  
Москва, Россия

Ключевые слова: корпус языка, морфологическая разметка, морфологический анализатор, словоизменение, машинно-ориентированная модель словоформы, система StarLing.

This paper describes development of a corpus of the Khakass language and design of the morphological parser for it. The work follows the framework of the RAS corporate project in regards to the development of corpora for languages of the Russian Federation, including Turkic minority languages such as the Khakass.

The algorithm of automatic morphological annotation is based on a dictionary taking into account phonetic alternation within the stem and on a computational model of wordform. The present work describes an attempt at creating such a model and defining a set of phonetic rules that constrain the choice of components of the wordform.

Key words: corpus of a language, morphological annotation, morphological parser, inflection, computational model of a wordform, StarLing system.

#### **1. О проекте создания хакасского корпуса**

В настоящее время в рамках Программы фундаментальных исследований РАН «Корпусная лингвистика. Создание и развитие корпусных ресурсов по языкам народов России»<sup>1</sup> (направление – создание и размещение в интернете корпусов текстов на тюркских языках<sup>2</sup> России, разработка корпусных технологий) ведется работа по созданию электронного корпуса хакасского языка.

Материалом для корпуса пока служат литературные тексты художественного жанра, оцифрованные и приведенные к стандартному формату, с переводом на русский язык. В распоряжении составителей корпуса есть также оцифрованная версия Большого хакасско-русского словаря на 22 тыс. слов под ред. О.В.Субраковой (далее – ХРС 2006) и иллюстративный материал к нему.

#### **2. Морфологическая разметка корпуса**

Морфологическая разметка текста состоит в приписывании словоформам информации об исходной форме слова и о совокупности их грамматических признаков.

---

<sup>1</sup> Подробнее см. <http://corpling-ran.ru/n3.html>

<sup>2</sup> Далее – ТЯ.

Эта информация включает следующие группы помет:

1. Основа, которой принадлежит словоформа (указывается «словарный вид» лексемы, при необходимости – часть речи<sup>3</sup>);
2. Словоизменительные характеристики (напр., падеж существительного, время глагола);
3. Информация о нестандартности грамматической формы, орфографических искажениях и т. п. [см. Ляшевская и др. 2005, 114].

### **3. Характеристики языка, релевантные для морфологической разметки**

Типологически хакасский язык принадлежит к языкам агглютинативного типа и обладает характерными для них признаками, как то:

- развитая система грамматически однозначных словоизменительных аффиксов, т.е. один аффикс выражает один грамматический признак;
- Отсутствие различных парадигматических классов в рамках одного словоизменительного типа – в отличие от флективных языков;
- отсутствие значимых чередований в основах, четкая фонетическая обусловленность использования алломорфов.

То есть, к основе в строгом порядке присоединяются однозначные аффиксы, границы морфем отчетливы, фонетические изменения на стыках морфем подчиняются строгим правилам. Однако эти конструктивные достоинства агглютинативного языка компенсируются количеством межморфемных сандхи, а множество грамматически однозначных аффиксов ведет к тому, что, например, парадигма имени состоит более чем из 600 мест. Не все формы одинаково частотны, но это – разрешенные для хакасского цепочки морфем, которые анализатор обязан учитывать.

### **4. Основные компоненты морфологического анализатора**

- словарь языка (словарь основ, учитывающий чередования);
- модель словоформы, ориентированная на автоматический анализ языка и опирающаяся на соответствующее грамматическое описание;
- набор фонетических правил, ограничивающих выбор алломорфов.

---

<sup>3</sup> Об относительности выделения частей речи в хакасском см. ниже.

#### 4.1. Словарь основ

Словарь основ представляет собой размеченную базу данных, содержащую слова в начальной форме (леммы) и не восстанавливаемые из начальной формы варианты чередований. Словарь извлечен из [ХРС 2006] с использованием технологий системы управления базами данных StarLing<sup>4</sup>. Импорт текстового файла словаря в StarLing и поэтапное создание на его основе многоуровневой лексико-грамматической базы данных описаны в работе [Крылов 2008]. С использованием технологий той же системы конвертирован в базу данных инвентарь морфем хакасского языка.

#### 4.2 Модель хакасской словоформы

Алгоритм автоматической морфологической разметки опирается на формальную машинно-ориентированную модель хакасской словоформы.

При построении модели мы использовали хакасские грамматики Баскакова [1953], [ГХЯ 1975], дополненные некоторыми инструментами грамматики порядков, традиционно применяющейся при описании агглютинативных языков; см. [Gleason 1955], [Мальцева 2004], [Володин, Храковский 1975], [Ревзин, Юлдашева 1969].

В рамках нашей модели словоформа представляет собой основу, к которой в фиксированной последовательности присоединяются словоизменительные аффиксы.

Морфологическая разметка содержит информацию о словоизменительных, но не о словообразовательных признаках лексемы. Т.е. парсер вычленяет в словоформе лишь словоизм. аффиксы (падеж, лицо, число, время, наклонение и т.п.) и пренебрегает теми, которые выполняют деривационную функцию и записаны в словаре основ. Например, в слове *палыхчыларыбыстың* 'наших рыбаков' словообразовательным является афф. *-чы* (аффикс деятеля, образующий слово 'рыбак' от *палых* 'рыба'), он рассматривается как часть основы и не учитывается при морф. анализе. Программа проанализирует слово как

*палыхчы-лар-ыбыс-тың*  
рыбак-Pl-Poss.1pl.-Gen

<sup>4</sup> Подробнее см. <http://starling.rinet.ru/program>

Для построения всех допустимых цепочек аффиксов из множества служебных морфем хакасского языка был выделен набор словоизменяемых показателей, выражающих все грамматические категории, существующие в языке. Каждому аффиксу и его алломорфам приписано грамматическое значение (напр., *лар/лер/нар/нер/ тар/тер* – афф. мн. числа, *да/де/та/те* – афф. местного падежа, *бын/бін/ пын/пін* – афф. 1.л. ед.ч. ...).

Ниже приведены лишь первые строки базы аффиксов в качестве иллюстрации модели хакасской словоформы (см. табл. 1).

#### 4.3. О выделении словоизменяемых классов в хакасской морфологии

Согласно сложившейся в алтаистике грамматической традиции ([Баскаков 1953], [ГХЯ 1975]) в хакасском языке выделяются три основных грамматических класса: имена, глаголы и неизменяемые (частицы, послелого, союзы и т.п.). Однако даже в вышеназванных источниках указывается, что дифференциация между грамматическими классами выражена слабо, особенно между разрядами имени [ГХЯ 1975, 82; Баскаков 1953, 403 и след.]: слово может трактоваться как существительное, прилагательное или наречие в зависимости от его синтаксической функции: существительное может выступать в роли определения: *тас туралар* ‘каменные дома’; прилагательное может выполнять в предложении любую функцию: *улуға орын пир* ‘старшему место уступи’.

В процессе работы над алгоритмом морф. разметки становится очевидна зыбкость границ между словоизм. классами не только в пределах имени, но также между именем, глаголом и т.н. «неизменяемым».

Имя в ТЯ присоединяет глагольные показатели: изменяется по лицам, принимая в зависимости от синтаксических функций 2 комплекта лично-числовых аффиксов – принадлежности (*минің хол-ым* ‘моя рука’, *минің алған-ым* ‘мое взятие’ -ым – афф. принадл. 1 л. ед.ч.) и лица (в составе именного сказуемого: *піс хакаспыс* ‘мы хакасы’, -пыс – афф. 1 л. мн.ч.).

Прилагательное в атрибутивной функции попадает в класс неизменяемых (*тас туралар* ‘каменные дома’), а будучи актантом принимает именные показатели: *кічіглер ойнапчалар* ‘маленькие (малыши) играют’.

В то же время хакасские глаголы, как и глаголы в большинстве языков

других групп, имеют формы, принимающие именные показатели – причастия, изменяющиеся по падежам и получающие показатели принадлежности; а большая часть финитных форм представляет собой те же причастия с показателями лица, теми же, что при именном сказуемом. Ср. также т.н. «алтайский тип сложноподчиненного предложения», где вторичная предикация выражается падежными формами причастия:

*хайди тоғын-ып үgren-гле-п-четкен-нер-ін чоохты-п пир-еңер*  
 как работать-Conv1 учиться (Refl)-Distr-Form-Prs.Pt-Pl-Poss3-Acc говорить-Conv1 дать-Imp.2pl  
 Расскажите, как они работают, учатся? букв. «Расскажите **учащихся-их** работая» [Мальцева 2004].

Таким образом, слова, традиционно относимые к разным частям речи – именам и глаголам, – могут принимать аффиксы одних и тех же грам. категорий: лица, числа, падежа, принадлежности. А это дает программе формальное основание причислять их к одному словоизменительному классу.

Принимая во внимание дискуссионность вопроса о выделении частей речи в ТЯ, а главное, нерелевантность такого выделения на этапе морфологической разметки, мы сочли целесообразным объединить две модели именных и глагольных словоформ в одну (табл. 1), где после основы слова в определенном порядке следуют грамматические показатели.

Заключение о принадлежности словоформы к определенной части речи в каждом конкретном случае становится актуальным уже на этапе синтаксической разметки.

Таблица 1

Модель изменяемой хакасской словоформы<sup>5</sup>

№	0 R (S)	1 Distr	2 Form	3 Emph	4 Part / Prosp	5 Dur	6 Neg	7 Tense (Pres Past, Future, Conv), Mood	8 Irr	9 Comit	10 Num (Pl)	11 Poss	12 APos	13 Case		14 Attr	15 Emph	16 Person (1, 2)	17 Adv
														Simple declension	Possessive declension				
1.		<i>гла</i>	<i>ып</i>	<i>даа</i>	<i>ыбыс</i>	<i>ча</i>	<i>ба</i>	Pres <i>ча</i>	<i>чы</i>	<i>лы</i>	<i>лар</i>	1sg <i>м</i>	<i>ни</i>	Gen <i>ның</i>	Gen <i>ның</i>	<i>хы</i>	<i>ох</i>	1sg <i>мын</i>	Manner <i>ни</i> ...
...		...	...	...	...	...	...	...	<i>X</i>	<i>F</i>	...	...	...	...	...	...	...	...	...

#### Необходимые пояснения к таблице

По горизонтали в верхней строке таблицы расположен перечень грамматических категорий, выраженных морфемами, следующими за корнем (основой), ячейки каждого столбца содержат аффиксы соответствующей грам. категории.

<sup>5</sup> За недостатком места здесь приводятся только верхние строки таблицы.

Одну позицию по горизонтали могут занимать морфемы одной или нескольких грам. категорий, каждая из которых (кроме R(S)) может не быть поверхностно выраженной, напр.: нулем выражается ед. число, им. падеж имени, показатель 3 л. ед. и мн. числа глаголов и имен).

### **Условные обозначения**

R(S) – корень или основа. Основа включает корень со словообразовательными показателями, присутствующий в словаре в качестве заголовка словарной статьи. Регулярное наличие показателя в словаре в составе заглавного слова служит критерием его включения в разряд словообразовательных.

Distr – дистрибутив, обозначает множественность субъекта или объекта действия; может также выражать многократность самого действия, итератив.

Form – формообразующий аффикс

Emph – эмфатическая частица

Perf/Prosp – перефектив (завершенность, законченность), проспектив (состояние, предшествующее действию)

Dur – дуратив (длительность действия)

Neg – отрицание

Tense, Mood – время, наклонение

Pres – настоящее время

Past – прошедшее время

Future – будущее время

Conv – деепричастие

Irr – ирреалис (гипотетичность)

Comit – комитатив, показатель совместности.

Num (Sg, Pl, Dual) – число

Poss – принадлежность

APos – показатель посессивного прилагательного. После этого аффикса субстантив принимает аффиксы падежей из посессивного склонения.

Case – падеж

Simple declension – набор падежных аффиксов простого склонения

Possessive declension – набор падежных аффиксов притяжательного склонения (после показателя принадлежности)

Gen – генетив (родительный падеж)

Attr – атрибутивный показатель

Adv – показатель наречия

Person – лицо

## **5. Фонетические закономерности, релевантные для морфологической разметки<sup>6</sup>**

Ниже приводится лишь несколько основных правил.

1. Важной фонетической закономерностью в хакасском является **сингармонизм**, действующий в большинстве агглютинативных языков, т.е. уподобление гласных внутри слова по фонетическим признакам ряда и огубленности.

В хакасском языке все гласные слова уподоблены по ряду первому гласному корня и могут быть либо заднерядными (*a, ы, у, о*), либо переднерядными (*i(u), e, ё, ö*), что сопровождается **аккомодацией согласных**

<sup>6</sup> Здесь рассматриваются только те закономерности, которые проявляются орфографически, т.к. морфологический анализатор работает только с письменными текстами.

(увулярные *х, ғ* при задних, велярные *к, ɟ* при передних): *харах-тар-ыбыс-та* ‘в наших глазах’ *кирек-тен-деңер* ‘по делам’.

2. **Ассимиляция согласных** (после глухих – глухие, после звонких – звонкие). Прогрессивная ассимиляция (уподобление последующего согласного предыдущим) ярко проявляется в вариативности звуков *л/н/т* в аффиксах мн.ч. (см. табл. 2);

3. **Озвончение глухих согласных *к, т, с, п, х* в интервокальной позиции**: *ат* ‘имя’ – *ады* ‘его имя’; *сас* ‘волос’ – *сазым* ‘мой волос’; *тап* ‘находить’ – *табыл-* ‘быть найденным; найтись’;

4. **Выпадение согласного в интервокальной позиции**: *суғ* ‘вода’ – *суу* (< *суғу*) ‘его вода’; *көг* ‘песня’ – *көө* ‘его песня’ (< *көгі*); *чілің +і > чілии* ‘мозг’; *тап + -ып > таап*;

5. **Выпадение узких гласных (*у, ы, і*) многосложных основ в позиции между согл. *р–н, л–н, й–н* перед афф. принадлежности *-ы, -і***: *пурун* ‘нос’ – *пурны* < *пуруны* ‘его нос’.

Сингармонизм и ассимиляция определяют правила выбора алломорфов; после того как словоформа скомпонована, на границах морфем начинают действовать внутренние сандхи. Окончательный фонетический облик словоформы является результатом последовательного применения этих правил.

Опираясь на вышеописанные закономерности, мы сформулировали правила обоих типов для последующего их применения при автоматическом морф. анализе. Ниже представлены некоторые их примеры.

Правила первой группы – **правила выбора** одного из нескольких фонетических вариантов аффикса – обобщают данные о сочетаемости алломорфов с предшествующим формантом. Примеры см. в табл. 2.

Аналогичные правила сформулированы для всех алломорфов хакасских словоизменяющих аффиксов, выделенных в табл. 1.

Правила второй группы – **поверхностные** – описывают механизм действия внутренних сандхи, например:

- озвончение глухих в интервокальной позиции (*ат + ы > ады*);
- выпадение согласного в интервокальной позиции и стяжение двух

обрамлявших его гласных в один долгий (*харах + ым > хараам, уйғу +-ға > уйғаа, маң +-ы > маа*);

- выпадение согласного *-ғ, -г, -ң* при стчении нескольких согласных (*суғ+-ға > суғ-а, көг+-ге > көг-е*);
- выпадение узких гласных (*у, ы, і*) многосложных основ на согласную перед аффт. принадлежности *-ы, -і* (*пурун 'нос' – пурны < пуруны 'его нос'*).

В настоящем виде модель хакасской словоформы и предложенные правила выбора и фонетических преобразований алломорфов должны применяться к широкому корпусу письменных текстов на хакасском языке и будут изменяться и дополняться в процессе работы.

Таблица 2

## Правила выбора аффиксов множественного числа

№	Фонетические свойства предшествующего элемента словоформы	Plur
1.	а) В элементах, предшествующих аффиксу, последний из гласных заднеряден ( <i>а, ы, о, у</i> ) & б) предшествующий элемент оканчивается на гласный или на звонкий неносовой согласный ( <i>б, в, ғ, д, ж, й, з, л, р</i> )	<i>лар</i> ( <i>тағ-лар</i> )
2.	а) В элементах, предшествующих аффиксу, последний из гласных переднеряден ( <i>-е, -и, -і, ө, ү</i> ) & б) предшествующий элемент оканчивается на гласный или на звонкий неносовой согласный ( <i>б, в, г, д, ж, й, з, л, р</i> )	<i>лер</i> ( <i>кізі-лер</i> )
3.	а) В элементах, предшествующих аффиксу, последний из гласных заднеряден ( <i>а, ы, о, у</i> ) & б) предшествующий элемент оканчивается на глухой согласный ( <i>п, ф, х, т, ш, с, ц, ч, щ</i> или оглушающийся звонкий <i>б, в, д, ж, з</i> в заимствованиях из русского)	<i>тар</i> ( <i>хус-тар, завод-тар</i> )
4.	а) В элементах, предшествующих аффиксу, последний из гласных переднеряден ( <i>-е, -и, -і, ө, ү</i> ) & б) предшествующий элемент оканчивается на глухой согласный ( <i>п, ф, к, т, ш, с, ц, ч, щ</i> или оглушающийся звонкий <i>б, в, д, ж, з</i> в заимствованиях из русского)	<i>тер</i> ( <i>түк-тер</i> )
5.	а) В элементах, предшествующих аффиксу, последний из гласных заднеряден ( <i>а, ы, о, у</i> ) & б) предшествующий элемент оканчивается на носовой согласный ( <i>м, н, ң</i> )	<i>нар</i> ( <i>хум-нар</i> )
6.	а) В элементах, предшествующих аффиксу, последний гласных переднеряден ( <i>-е, -и, -і, ө, ү</i> ) & б) предшествующий элемент оканчивается на носовой согласный ( <i>м, н, ң</i> )	<i>нер</i> ( <i>күн-нер</i> )

## Литература

Баскаков 1953 – Хакасско-русский словарь / Под ред. Н.А.Баскакова, с приложением грамматического очерка хакасского языка Н.А.Баскакова. М., 1953. С. 361–485.

Володин, Храковский 1975 – Володин А.П., Храковский В.С. Типология морфологических классификаций глагола (на материале агглютина-

тивных языков) // Типология грамматических категорий: Мещаниновские чтения. М.: Наука.

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

Крылов 2008 – *Крылов С.А.* Стратегии применения интегрированной информационной среды StarLing в корпусной лингвистике и в компьютерной лексикографии // *Orientalia et classica*. Труды Института восточных культур и античности. Выпуск XIX. Аспекты компаративистики. 3. М., РГГУ, 2008. С. 649–668.

Крылов 2011 – *Крылов С.А.* Использование системы StarLing при создании морфологически аннотированного корпуса современного монгольского языка. – *на правах рукописи*.

Ляшевская и др. 2005 – *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. 111—135.

Мальцева 2004 – *Мальцева В.С.* Структура глагольной словоформы в сагайском диалекте хакасского языка (говор с. Казановка). М., 2004.

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

Ревзин, Юлдашева 1969 – *Ревзин И.И., Юлдашева Г.Д.* Грамматика порядков и ее использование.// *Вопросы языкознания*, №1. с. 42–56.

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

ХРС 2006 – Большой хакасско-русский словарь / Под ред. О.В.Субраковой. Новосибирск, «Наука», 2006.

*Çağrı Çöltekin* A Freely Available Morphological Analyzer for Turkish / <http://www.let.rug.nl/coltekin/papers/coltekin-lrec2010.pdf>;  
<http://www.let.rug.nl/coltekin/trmorph/>

Gleason 1955 – *Gleason, H.* Introduction to descriptive linguistics. New York, 1955.