МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12–06–06061–г),
гранта РГНФ (проект № 12–04–14154–г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:
В. А. Баранов, д-р филол. наук, проф.
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

И741 **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

К МОРФОЛОГИЧЕСКОЙ РАЗМЕТКЕ КОРПУСА ХАКАССКОГО ЯЗЫКА

А. В. Шеймович

Институт языкознания РАН, Москва

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.

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»1 (направление — создание и размещение в интернете корпусов текстов на тюркских языках2 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 [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.

Эта информация включает следующие группы помет:

1. Основа, которой принадлежит словоформа (указывается «словарный вид» лексемы, при необходимости — часть речи3);

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 раз-метки

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.

4.1. Словарь основ

Словарь основ представляет собой размеченную базу данных, содержащую слова в начальной форме (леммы) и не восстановимые из начальной формы варианты чередований. Словарь извлечен из [ХРС, 2006] с использованием технологий системы управления базами данных StarLing1. 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 [Баскаков, 1953; ГХЯ, 1975], дополненные некоторыми инструментами грамматики порядков, традиционно применяющейся при описании агглютинативных языков; см. [Gleason, 1955; Мальцева, 2004; Володин и Храковский, 1975; Ревзин и Юлдашева, 1969].

---

[1] Подробнее см. http://corpling-ran.ru/n3.html

[2] Далее – ТЯ.

[3] Об относительности выделения частей речи в хакасском см. ниже.

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 словоизменительные аффиксы (падеж, лицо, число, время, наклонение и т.п.) и пренебрегает теми, которые выполняют деривационную функцию и записаны в словаре основ. Например, в слове *палыхчылarыбыстың* ‘наших рыбаков’ словообразовательным является аффикс *-чы* (аффикс деятеля, образующий слово ‘рыбак’ от *палых* ‘рыба’), он рассматривается как часть основы и не учитывается при морфологическом анализе. Программа проанализирует слово как

палыхчы-лар-ыбыс-тың

рыбак-Pl-Poss.1pl.-Gen

Для построения всех допустимых цепочек аффиксов из множества служебных морфем хакасского языка был выделен набор словоизменительных 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. таб. 1).

4.3. О выделении словоизменительных 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, послелоги, 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 за-висимости от его синтаксической функции: существительное может высту-пать в роли определения: тас туралар ‘каменные дома’; прилагательное может выполнять в предложении любую функцию: улуға орын пир ‘старше-му место уступи’.

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

---

[1] Подробнее см. http://starling.rinet.ru/program

Имя в ТЯ присоединяет глагольные показатели: изменяется по лицам, принимая в зависимости от синтаксических функций 2 комплекта лично-числовых аффиксов — принадлежности (*миніҥ хол-ым* ‘моя рука’, *миніҥ 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 неизменяе-мых (*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 — причас-тия, изменяющиеся по падежам и получающие показатели принадлежности; а большая часть финитных форм представляет собой те же причастия с пока-зателями лица, теми же, что при именном сказуемом. Ср. также т.н. «алтай-ский тип сложноподчиненного предложения», где вторичная предикация вы-ражается падежными формами причастия:

хайди тоғын-ып ***ӱгрен-гле-п-четкен-нер-ін*** *чоохты-п пир-еҥер*

как работать-Conv1 учиться (Refl)-Distr-Form-Prs.Pt-Pl-Poss3-Acc гово-рить-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 граммати-ческих категорий: лица, числа, падежа, принадлежности. А это дает програм-ме 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 морфоло-гической разметки, мы сочли целесообразным объединить две модели имен-ных и глагольных словоформ в одну (табл. 1), где после основы слова в оп-ределенном порядке следуют грамматические показатели.

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

Табл. 1. Модель изменяемой хакасской словоформы[1]

№	0	1	2	3	4	5	6	7	8	9	10	11	12	13		14	15	16	17
	R (S)	Distr	Form	Emph	Perf / Prosp	Dur	Neg	Tense (Pres Past, Future, Conv), Mood	Irr	Comit	Num (	Poss	APos	Case		Attr	Emph	Person (1	Adv
														Simple declen-sion	Poss es				
1. …		ғла …	ып …	даа …	ыбыс …	ча …	ба …	Pres ча …	?ых …	лығ …	лар …	1sg м …	ни …	Gen ның …	Gen ның …	хы …	ох …	1sg мын …	Manner ни …

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 соответсвующей грамматической категрии.

Одну позицию по горизонтали могут занимать морфемы одной или нескольких грамматических 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 — перефктив (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 субстанитв принимает аффиксы падежей из посессивного склонения.

Case — падеж

Simple declendion — набор падежных аффиксов простого склонение

Possessive declendion — 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[1]

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 (*а, ы, у, о*), либо переднерядными (*і(и), е, ӱ, ӧ*), что сопровождается **аккомодацией согласных** (увулярные *х, ғ* при задних, велярные *к, г* 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 морфологическом 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 словоизменительных аффиксов, выделенных в табл. 1.

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

---

[1] Здесь рассматриваются только те закономерности, которые проявляются орфографически, т.к. морфологический анализатор работает только с письменными текстами.

– 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 стечении нескольких согласных (суғ+-ға > суғ-а, көг+-ге > көг-е);

– выпадение узких гласных (*у, ы, і*) многосложных основ на согласную перед афф. 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.

Табл. 2. Правила выбора аффиксов множественного числа

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

Список литературы

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

Володин и Храковский, 1975 — Володин А.П., Храковский В.С. Типология морфологических классификаций глагола (на материале агглютинативных языков) // Типология грамматических категорий: Мещаниновские чтения. М.: Наука, 1975

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.

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. С. 111–135.

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

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

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

ХРС, 2006 — Большой хакасско-русский словарь / Под ред. О.В. Субраковой. Новосибирск, 2006.

Çöltekin, 2010 — Çağrı Çöltekin. A Freely Available Morphological Analyzer for Turkish // Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta, May 2010. http://www.let.rug.nl/coltekin/papers/coltekin-lrec2010.pdf

Gleason, 1955 — Gleason, H. Introduction to descriptive linguistics. New York, 1955.