Аппарат автоматического морфологического анализа для корпуса хакасского языка[1]

A morphological parser for the Khakass Language Corpus

Дыбо А.В., Шеймович А.В.
Dybo A.V., Sheimovich A.M.

В статье освещается работа по созданию корпуса текстов хакасского языка и снабжению его научным аппаратом, а именно — инструментами морфологического анализа. Основное внимание уделяется созданию компьютерно ориентированной модели хакасской словоформы. Работа ведется в рамках масштабного проекта по созданию корпусных ресурсов языков народов России, в частности корпусов на малых тюркских языках РФ. Алгоритм автоматического морфологического аннотирования базируется на словаре, учитывающем фонетическое чередование внутри основы, и на компьютерной модели словоформы. В настоящей работе представлена попытка создания такой модели и определения набора фонетических правил, которые ограничивают выбор компонентов словоформы.

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

This paper describes the Russian Academy of Science's project to develop a corpus of the Khakass language and to design a morphological parser for this corpus, carried out within a broader corpora development macroproject for the minority languages of Russia, including the Turkic languages. The algorithm for automatically annotating Khakass

[1] Работа выполнена в рамках проекта РГНФ № 15-04-12030 «Система автоматического морфологического и синтаксического анализа для корпусов миноритарных тюркских языков России».

morphology is based on dictionary entries that take into account phonetic alternations within the word stem and on computational modeling of Khakass word forms (lexical templates). The present work describes the attempt to create such a model and to define a set of phonological rules that constrain the choice of what components can constitute a Khakass word.

Keywords: Khakass, corpus linguistics, morphological annotation, automatic parser, derivational and inflectional morphology, computational modeling of lexical morphology, StarLing database system

1. О корпусе хакасского языка

В настоящее время в рамках 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) ведется работа по развитию, пополнению и аннотированию электронного корпуса хакасского языка. Корпус состоит из диалектного подкорпуса и корпуса литературных и публицистических текстов, снабженных русским переводом. В распоряжении составителей корпуса есть также оцифрованная версия Большого хакасско-русского словаря на 22 тыс. слов под ред. О.В. Субраковой (далее — БХРС) [БХРС 2006] и иллюстративный материал к нему. Тексты корпуса планируется снабдить морфологической, словообразовательной, семантической и синтаксической разметкой. В данной статье рассматривается разработка автоматического анализатора для морфологической разметки хакасских текстов.

1.1. Морфологическая разметка корпуса

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). Работа была начата с морфологического аннотирования

(разметки). Разметка выполняется автоматически специальной программой — морфологическим парсером. Морфологическая разметка текста состоит в приписывании каждой словоформе информации о словарном вхождении (словарный вид лексемы, часть речи) и о грамматических признаках (напр., падеж существительного, время глагола и т.п.).

2. Компоненты морфологического анализатора

Аппарат автоматического морфологического анализа состоит из следующих компонентов:

- словарь языка (словарь основ, учитывающий чередования);
- грамматика, ориентированная на автоматический анализ языка (инвентарь морфем, порядок их следования и правила сочетания).

2.1. Словарь

Словарь основ автоматическим образом извлечен из БХРС с использованием системы управления базами данных StarLing.

После обработки файла словаря системой StarLing словарь представляет собой размеченную базу данных, содержащую полнозначные слова (основы) в начальной форме — леммы — и не восстановимые из начальной формы варианты чередований. Если начальная форма состоит не только из корневой морфемы, но включает в себя также словообразовательные показатели, эти показатели и их значения фиксируются в специальных полях базы данных для их последующего вывода в словообразовательной разметке корпуса.

2.1.1. Необходимые замечания относительно словаря основ

Организация статьи словарной базы.

Верхнее поле базы **FIELD1** дублирует содержание словарной статьи БХРС, сохраняя ее формат, шрифтовые выделения,

разделители и весь иллюстративный материал, являющийся ценной составляющей корпуса.

В поле **HEADWORD** выписано заголовочное слово словарной статьи.

В поле **ALTERNAT** автоматически скопированы основы глаголов, которые в БХРС выписаны в слэшах после формы инфинитива (Dat от прич. на *-Ар*). Основа хакасского глагола в общем случае не может быть автоматически получена из формы инфинитива путем отсечения -АрГа: если основа оканчивается на глухую согласную морфонему, то эта согласная перед показателем -АрГа озвончена в интервокальной позиции. Если основа оканчивается на шумную звонкую согласную морфонему Г, то эта согласная перед показателем -АрГа в стандартном случае выпадает, а гласная в показателе принимает стяженную форму в зависимости от гласной корня. Только в случае, если основа оканчивается на сонорную согласную морфонему, ее облик перед показателем инфинитива сохраняется. Если основа глагола оканчивается на гласную морфонему, то происходит стяжение этой гласной с вокалическим началом аффикса (см. об этом типе стяжения ниже). Ср.:

ААЛЛАДАРҒА /ааллат-/ принимать гостей
ПООРҒА /поғ-/ перетягивать, стягивать *что-л*
ПОРАНАРҒА /поран-/ жить в беспорядке
АҒЫННИРҒА /ағынна-/ улетать на зимовку.

В поле **ALTERNATEN** выписана вручную словоизменительная именная основа в том случае, если она не совпадает формально с начальной словарной формой имени и это несовпадение не является результатом действия регулярного правила. Именно эта основа используется при выборе морфонологического варианта добавляемого к ней аффикса. Эта основа не может быть получена из содержания словарной статьи БХРС автоматически. В БХРС при именных основах выписаны в скобках некоторые возможные чередования, представленные в словоизменении, но, если, по мнению

авторов словаря, основа изменяется только по регулярным правилам, они могут не выписываться. В результате выписывание чередований именной основы в БХРС проведено крайне непоследовательно. Так, в хакасском есть регулярное правило об озвончении глухих согласных в интервокальной позиции; вот как выглядят две словарные статьи на одной странице:

ДЕКАНАТ деканат; математика факультедінің деканады деканат факультета.

ДЕЛЕГАТ (-ды) делегат // делегатский; съезд делегаттары делегаты съезда.

Из примера видно, что оба эти слова ведут себя одинаково в отношении озвончения в интервокальной позиции, однако их подача в словарных статьях различается. Так как правило об озвончении действует регулярно, то в статьях для обоих указанных слов поле ALTERNATEN остается незаполненным.

Нерегулярные же чередования типа **ПОЛКОВОДЕЦ (-зі)** отражены в поле ALTERNATEN (полководес). Эта форма выступает перед аффиксами, начинающимися на гласную морфонему, словарная форма — в абсолютном конце слова или перед аффиксами, начинающимися на согласную морфонему.

Для имен, содержащих беглые гласные (узкие *у, ы, і* в закрытом конечном слоге основы, находящиеся между согл. *р–н, л–н, й–н*), в поле ALTERNATEN также выписаны чередующиеся основы:

ОРЫН (-ны) место: ALTERNATEN орн.

Говоря о поле ALTERNATEN, отдельно следует упомянуть заимствованные из русского языка существительные с основой на оглушающийся звонкий *б, в, д, ж, з* (*завод, скандинав*). Из выбора алломорфов при словоизменении видно, что словоизменительная основа этих слов заканчивается

на глухую морфонему *п, т, ш, с*, но пишется во всех сочетаниях со звонким. Поэтому словарная (орфографическая) основа подобных слов совпадает с начальной формой, отраженной в поле FIELD1 (ЗАВОД), а реальная со знаком * выписывается в поле ALTERNATEN (*завот**) — так что аффиксы после этой основы выбираются в том варианте, который должен следовать после глухой согласной: Loc *заводта*, Pl *заводтар*.

В поле **DERIV** записана основа слова, расчлененная на словообразовательные элементы. Границы между элементами обозначены знаком =; между компонентами композита — знаком +.

Поле **DERIVGLOSS** содержит словообразовательную глоссировку словоформы. При глоссировке используются те же знаки, внутри кумуляции ставится точка (“.”).

В поле **SEMGLOSS** записан перевод основы, который должен фигурировать в глоссировках. Это более краткий, лаконичный вариант перевода, не содержащий описательной части, включенной в основное поле **FIELD 1**. Так, в статье: **ХЫЙҒЫС** I крюк для обработки овчины, поле **SEMGLOSS** содержит только слово «крюк». Поле заполнялось в два прохода: автоматически в него выписывалось первое слово из переводной части статьи, затем краткий перевод редактировался вручную.

В поле **SEMTAG** записана семантическая информация о лексеме в виде набора семантических помет.

Поле **ETYM**, предназначенное для этимологических помет, на настоящий момент заполнено только для русских заимствований и слов, пришедших в хакасский через русское посредство (для тюркских по происхождению слов впоследствии планируется сделать ссылки на тюркскую этимологическую базу).

Поле **REST** содержит значения слова и примеры словоупотребления. См. рис. 1–4.

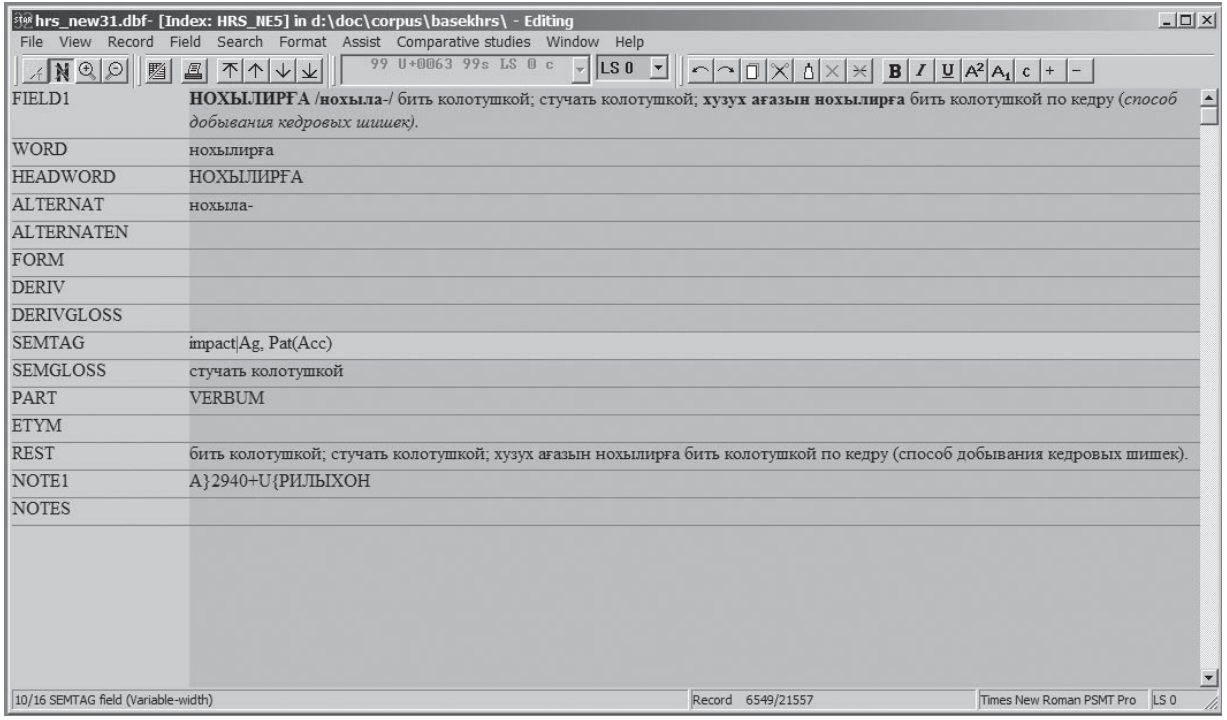
hrs_new31.dbf - [Index: HRS_NE13] in d:\doc\corpus\basekhrs\ - Editing

File View Record Field Search Format Assist Comparative studies Window Help

FIELD1	**ЗАВОД** I завод // заводской, заводский; **ат заводы** конный завод; **вагон пӱдӟен завод** вагоностроительный завод; **палых заводы** рыболовный завод; **халас заводы** хлебозавод; **завод лабораториязы** заводская лаборатория; **заводтын устаг-пастаа** заводское управление.
WORD	завод
HEADWORD	ЗАВОД
ALTERNAT	
ALTERNATEN	завот
FORM	
DERIV	
DERIVGLOSS	
SEMTAG	org=constr
SEMGLOSS	завод
PART	NOMEN
ETYM	rus
REST	завод // заводской, заводский; ат заводы конный завод; вагон пӱдӟен завод вагоностроительный завод; палых заводы рыболовный завод; халас заводы хлебозавод; завод лабораториязы заводская лаборатория; заводтын устаг-пастаа заводское управление.
NOTE1	ДОВАЗ
NOTES	

10/16 SEMTAG field (Variable-width) Record 2200/21557 Times New Roman PSMT Pro LS 0

Рис. 1



Стр. 16 — рис. 2. Стр. 17 — рис. 3

hrs_new31.dbf - [Index: HRS_NE5] in d:\doc\corpus\basekhrs\ - Editing

File View Record Field Search Format Assist Comparative studies Window Help

FIELD1	**ОРЫН (-ны) I** 1) место; пространство; **истiг орын** удобное место; **кiзi орны** чужое место; **орын чоғыл** нет места; **от орны** очаг, место для огня *(напр.,* в юрте); **тоғыс орны** место работы; **тыныҥан орын** место отдыха; **орынан даа хыймырабады** с места не встал; 2) местность, **аҕас чох орын** безлесная местность; **сiлiг орында чуртаҥчам** [я] живу в красивой местности; 3) должность, служба; **чахсы орын** хорошая должность; **орын чох халды** [он] остался без должности; **паска орынча парыбыстым** [я] ушёл на другую должность; 4) место, часть текста, произведения; **чоохтын ин хыныг орны** самое интересное место в рассказе; **пу орынны хатаҥ хыгырыбызарға** это место прочитать заново; 5) место *(отдельная вещь багажа, груза);* **минiҥ пiр орын** у меня одно место; **сұғ орны** русло реки; **а пала орны** *анат.* послед *(у женщины);* **пик орын** надёжное место; **тик орын** пустое место *(не соответствующее занимаемой должности);* **улуға орын пир, кiчiге полыс пир** *погов.* старшему место уступи, младшему помощь окажи; **орынан даа турбаан** *погов. соотв.* даже пальцем не пошевелил; **орын пiлiн** *погов.* знай своё место *(букв.* не вмешивайся в чужие дела).
WORD	орын
HEADWORD	ОРЫН
ALTERNAT	
ALTERNATEN	орн
FORM	
DERIV	
DERIVGLOSS	
SEMTAG	space
SEMGLOSS	место
PART	NOMEN
ETYM	
REST	1) место; пространство; истiг орын удобное место; кiзi орны чужое место; орын чоғыл нет места; от орны очаг, место для огня (напр., в юрте); тоғыс орны место работы; тыныҥан орын место отдыха; орынан даа хыймырабады с места не встал; 2) местность, аҕас чох орын безлесная местность; сiлiг орында чуртаҥчам [я] живу в красивой местности; 3) должность, служба; чахсы орын хорошая должность; орын чох халды [он] остался без должности; паска орынча парыбыстым [я] ушёл на другую должность; 4) место, часть текста, произведения; чоохтын ин хыныг орны самое интересное место в рассказе; пу орынны хатаҥ хыгырыбызарға это место прочитать заново; 5) место (отдельная вещь багажа, груза); минiҥ пiр орын у меня одно место; сұғ орны русло реки; а пала орны анат. послед (у женщины); пик орын надёжное место; тик орын пустое место (не соответствующее занимаемой должности); улуға орын пир, кiчiге полыс пир погов. старшему место уступи, младшему помощь окажи; орынан даа турбаан погов. соотв. даже пальцем не пошевелил; орынҥ пiлiн погов. знай своё место (букв. не вмешивайся в чужие дела); с притяж. афф. 3 л. [его] место; чурт орны место, где стоит, стоял или будет стоять дом.
NOTE1	НЫРО
NOTES	

10/16 SEMTAG field (variable-width) | Record 7203/21557 | Times New Roman PSMT Pro | LS 0

hrs_new31.dbf - in d:\doc\corpus\baselhrs\ - Editing
File View Record Field Search Format Assist Comparative studies Window Help
* Begin of text *
LS 0
Times New Roman PSMT Pro
10/16 SEMTAG field (variable-width)
Record 7/21557
FIELD1 ААЛДАГЫ сельский; ааЛдагы үтретчі сельский учитель; ааЛдагы чон сельское население, селяне.
WORD ааЛдагы
HEADWORD ААЛДАГЫ
ALTERNAT
ALTERNATEN
FORM
DERIV ААЛ=ДА=ГЫ
DERIVGLOSS село=Loc=Ar
SEMTAG space:setl
SEMGLOSS сельский
PART NOMEN
ETYM
REST сельский; ааЛдагы үтретчі сельский учитель; ааЛдагы чон сельское население, селяне.
NOTE1 Ы}2940+U{АДЛАА
NOTES

Рис. 4

2.1.2. О выделении грамматических классов и частей речи в хакасской морфологии

Согласно традиции, принятой в описании алтайских языков (см. [Баскаков 1953; ГХЯ 1975]), в хакасском языке выделяются три основных грамматических класса: имена, глаголы и неизменяемые (частицы, послелоги, союзы и т.п.).

В словарной базе данных помета «часть речи» проставлялась автоматически. Глаголам в словаре основ была присвоена частеречная помета Verbum. Она присваивалась словам, оканчивающимся на *-рға/-рге*, если этим буквосочетаниям предшествовало не менее двух гласных. Пометы для неизменяемых слов (частиц, союзов, междометий) были автоматически скопированы из БХРС, после этого оставшимся словам была присвоена помета Nomen (имя).

В хакасской грамматике традиционно принято деление имен на более мелкие разряды: существительное, прилагательное, наречие, числительное, местоимение. Однако при этом указывается, что дифференциация между грамматическими классами выражена слабо, особенно между именными разрядами — существительными, прилагательными и наречиями [ГХЯ 1975: 82; Баскаков 1953: 403 и след.]: 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 туралар* ‘каменные дома’; прилагательное может выполнять в предложении функцию существительного или наречия, ср. *улуға орын пир* ‘старшему место уступи’.

В процессе работы над алгоритмом автоматической морфологической разметки стали очевидными некоторые особенности тюркского словоизменения, которые обычно затушевываются при выделении частей речи традиционными тюркскими грамматиками, написанными по образцу классической грамматики. Анализ этих особенностей приводит

к выводу о зыбкости границ между частями речи не только в пределах имени, но также и между именем, глаголом и неизменяемыми частями речи.

Имя в хакасском языке (как и в других тюркских) может принимать грамматические категории, обычно считающиеся глагольными: изменяться по лицам, присоединяя в зависимости от синтаксических функций два комплекта лично-числовых показателей — личной принадлежности (*минің хол-ым* ‘моя рука’, *минің алған-ым* ‘мое взятие’ *-ым* – показатель 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 тура* ‘каменный дом’), а будучи актантом глагола присоединяет именные показатели: *кічіглер ойнапчалар* ‘маленькие (малыши) играют’.

В то же время тюркские и хакасские глаголы, как и глаголы в большинстве языков других групп, имеют формы с именными показателями — причастия, которые изменяются по падежам, а иногда принимают и показатели принадлежности. Ср. также так называемый «алтайский тип сложноподчиненного предложения», где вторичная предикация[2] выражается падежными формами причастия, например:

(1) *хайди тоғын-ып*
Как работать-Conv1
ӱгрен-гле-п-четкен-нер-ін
учиться (Refl)-Distr-Form-Prs.Pt-Pl-Poss3-Acc
чоохты-п *пир-еңер*
говорить-Conv1 дать-Imp.2pl
Расскажите, как они работают, учатся? *букв.*
‘Расскажите в настоящее время **учащихся-их** работая’

[Мальцева 2004]

[2] Вторичная предикация образуется неличными формами глагола — инфинитивом, причастием, деепричастием.

Более того, похожим образом иногда ведут себя слова, причисляемые к неизменяемым служебным частям речи. Слово *осхас*, определяемое словарем как «послелог, выражающий подобие: ‘также как; как будто, подобно’» в зависимости от объекта подобия может принимать как именные, так и глагольные показатели (при этом смысловое слово выступает обычно в начальной форме).

(2) *Мин осхас-тар-ға хатхыр-ча-лар*
я как-Pl-Dat смеяться-Pres-Pl
Смеются над такими, как я.

(3) *Көй пар-ған осхас-пын мин*
гореть идти-Past словно-1SG
Я словно вся выгорела.

Таким образом, слова, традиционно относимые к разным частям речи — именам, глаголам и неизменяемым, — могут принимать аффиксы одних и тех же грамматических категорий: лица, числа, падежа, принадлежности. При этом можно заметить, что набор грамматических категорий имени является подмножеством набора грамматических категорий глагола: в именной части модели словоформы остаются незаполненными позиции таких категорий, как время, наклонение, отрицание и еще нескольких, которые можно отнести к исключительно глагольным. Вопрос выделения частей речи для тюркских языков является дискуссионным, к тому же на этапе автоматической разметки он преждевременен, т. к. здесь в первую очередь важно определение словарной основы, а также служебных морфем с их грамматическими значениями. Поэтому выносить решение о принадлежности слова к части речи было решено в процессе синтаксической разметки, которая остается за пределами настоящей работы.

2.2. Грамматика (инвентарь морфем, порядок их следования и правила сочетания)

Алгоритм морфологического анализа опирается на формальную компьютерно ориентированную модель хакасской словоформы. С учетом всего сказанного о частях речи в хакасском языке, для удобства анализа была построена объединенная модель для именных и глагольных словоформ в виде таблицы. См. таблица 2, с. 32. При этом для именных основ остаются незаполненными позиции с 1 по 8-ю, а в позиции 16 (Person) отсутствуют аффиксы императива (Imp и Prec).

В рамках представленной ниже модели словоформа представляет собой основу, к которой в определенной последовательности присоединяются словоизменительные аффиксы. Основа состоит из корня, к которому могут присоединяться аффиксы словообразования.

Содержательными источниками для модели послужили грамматические описания хакасского языка, содержащиеся в [Баскаков 1953] и в [ГХЯ 1975], а также [БХРС 2006]. Формально же данная таблица укладывается в теоретическую модель грамматики порядков, введенную в оборот американским дескриптивистом Г. Глисоном и с тех пор традиционно применяющеюся при описании агглютинативных языков. В рамках данного подхода описывались тюркские языки, в частности хакасский [Мальцева 2004], и палеоазиатские языки. Ср. аналогичное описание корейской глагольной словоформы в [Martin 1992: 244–274] и попытку машинной реализации морфологического и синтаксического анализа корейского языка на основе такой модели в [Бречалова 2009].

Отличительной чертой нашего подхода к морфологической разметке является ***разграничение словообразовательной и словоизменительной морфологии***. Оно производится чисто механически на том основании, что основы со словообразовательными аффиксами присутствуют в словаре

в качестве заголовков словарных статей, часто с пометами, указывающими на наличие в слове словообразовательных аффиксов, напр.:

ПОЛЫНАРҒА *возвр. от* **поларға;** мочь, быть в силах *(в состоянии способным что-л. делать для себя)*;

АЙТЫРАРҒА понуд. от **айтарға; хабар айтырарға** просить передать известие;

ХОҢЫРОҶАХ *уменьш. от* **хоңыро** бубенчик.

Помимо этого существуют и другие причины того, что в нашей модели морфологического анализа не учитываются словообразовательные показатели:

1. Словообразовательные показатели при включении в разметку, нарушают важный принцип грамматики порядков, а именно принцип однократности появления в словоформе граммем одной категории. В качестве примера такого нарушения можно привести аффиксы залога в хакасском глаголе: Двойной каузатив:

(4) *ол паба-зы-на* *көнек* ***тур-ғыс-тыр-ча***
он отец-Poss3-Dat ведро стоять-Caus-Caus-Pres
Он заставляет отца поставить ведро.

Двойной пассив:

(5) *лампа* *чил-наң* *чайха-л-ыл-ча*
лампа ветер-Abl качать-Pass-Pass-Pres
Лампа качается от ветра.

Пассив+Рефлексив:

(6) *кізі* *чайха-л-ын-ча*
человек качать-Pass-Refl-Pres
Человек качается.

[Мальцева 2004]

Аффиксы конверсии (формообразующие аффиксы, обычно трактующиеся как аффиксы соответствующей части речи) также могут встречаться в слове более чем по одному разу

и в разном порядке, переводя слово из одного лексико-грамматического разряда в другой:

(7) *чӳк-те-н-ӷік* ‘котомка’
ноша- Oper[3]-Refl > *чӳк-те-н* ‘нести’ + *ӷік* (Dimin)

(8) *ачы* ‘киснуть’
ачы-ғ ‘горький’
киснуть-NomStat
ачы-ғ-ла-н-арға ‘скорбеть’
киснуть-NomStat-Oper-Refl-Infin

2. Фразеологизованность семантики производных лексем: значение производной словоформы не выводится из суммы значений корня и аффиксов, в силу этого включение словообразования в систему морфологического анализа не представляется значимым:

(9) *хас* ‘берег’ — *хаста-* ‘идти вдоль чего-л.’:
хана хастирға ‘идти вдоль забора’
суғ хастирға ‘идти по берегу’
хастирға < *хас+та+арға*
берег- Oper-Inf

По этим причинам на нынешнем этапе работы программа-парсер вычленяет в слове лишь аффиксы, имеющие грамматическое значение (падеж, лицо, число, время, наклонение, принадлежность и т. п.) и пренебрегает деривационными. Например, в слове *палыхчыларыбыстың* ‘наших рыбаков’ появляется афф. *-чы* (суффикс деятеля, образующий слово ‘рыбак’ от *палых* ‘рыба’), он рассматривается как часть основы и не учитывается при морф. анализе. Программа проанализирует слово как

[3] Oper — формант, выражающий лексическую функцию образования отыменного переходного глагола; ср. Func — формант, выражающий лексическую функцию образования отыменного непереходного глагола. См. [Жолковский, Мельчук 1967].

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

Однако иногда парсер анализирует и словообразовательные аффиксы, такие как, например, атрибутивный показатель (Attr) и показатель наречия образа действия (Adv). Attr (*ғы*) и Adv (*ли*) — показатели, которые функционируют обычно как формообразующие, прибавляясь к чистой основе; тогда соответствующая единица попадает в словарь (*амғы* ‘современный’, *хысти* ‘всю зиму’), и аффиксы не учитываются при морф. анализе. Однако они могут свободно присоединяться к концу длинной словоформы, содержащей уже словоизменительные аффиксы, при помещении ее в определенные синтаксические условия; тогда их нет в составе словарной словоформы (*цирк-та-ғы* < цирк-Loc-Attr ‘цирковой’, *хомды-да-ғы* < гроб-Loc-Attr ‘как в гробу’; *ас-ған-ни* < обгонять-Past-Adv ‘обгоняя’). В этом случае парсер подвергает эти аффиксы анализу[4].

[4] По мере работы с корпусом естественного языка и поступления новых языковых данных мы продолжаем редактировать автоматический анализатор. Коррекции подвергается также модель изменяемой словоформы, лежащая в его основе. Так, при обработке полевых материалов были обнаружены новые примеры рекурсии, побудившие нас создать в модели дополнительные слоты для дублирующих наборов числовых, посессивных и падежных показателей. Сделаем здесь следующее замечание. Аффиксы конверсии (такие как *ЛІГ, КІ*) в принципе в хакасском языке, как и в других тюркских, работают рекурсивно и теоретически допускают сколько угодно циклов с добавлением чисто словоизменительных аффиксов к новой «основе»; таким образом, можно было бы создать автомат, обрабатывающий каждую словоформу циклически. Однако а) случаи с количеством рекурсий больше двух практически в текстах не встречаются (а случаев с двойной рекурсией очень мало), б) такой автомат увеличит количество альтернативных разборов во много раз; вследствие этого мы сочли его нецелесообразным.

В конечном же варианте размеченного корпуса будет присутствовать словообразовательная разметка. В словаре основ она делается вручную.

Те же требования грамматики порядков (и автоматического анализа) – жесткий порядок следования и однократность появления аффиксов внутри словоформы – приводят к некоторым различиям в отображении грамматических показателей между нашей грамматикой и традиционными грамматиками хакасского языка. Так, некоторые показатели, традиционно трактуемые как один целостный аффикс, проанализированы нашим парсером как цепочка аффиксов. В качестве примера можно привести показатели длительности действия (Dur) *-чат* и прошедшего времени (Past) *-хан*, которые при последовательном употреблении считаются в ГХЯ в зависимости от синтаксической функции словоформы то аффиксом причастия настоящего времени: *хас-чатхан* ‘копающий’ [ГХЯ 1975: 232], то показателем прошедшего определенного времени: *узу-п-чатхан* ‘я спал (*тогда*)’ [ГХЯ 1975: 216]. В связи с этим мы рассчитываем, что в окончательной версии морфологического анализа парсер будет предъявлять пользователю 2 варианта морфологического разбора, один из которых трактовал бы выделяемые показатели в терминах традиционной грамматики.

2.3. Правила выбора фонетических вариантов аффиксов

На работу морфологического анализатора влияют также фонетические закономерности, действующие в языке[5]. При сочетании основы и словоизменительного

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

аффикса, а также самих аффиксов друг с другом их фонетический облик может изменяться в результате т.н. внутренних сандхи[6], обусловленных сингармонизмом, аккомодацией, ассимиляцией и другими фонетическими процессами.

Законы сингармонизма, ассимиляции и аккомодации определяют ***правила выбора*** алломорфов в каждой из возможных в хакасском языке цепочек аффиксов[7]. Эти правила описывают изменения, возникающие вследствие взаимодействия между морфемами: между основой и аффиксами, между соседними аффиксами. После этого вступают в действие ***внутренние сандхи*** на границах морфем (интервокальные озвончения, выпадения и стяжения — т.н. «поверхностные правила»). Окончательный фонетический облик словоформы является результатом последовательного применения этих двух видов правил.

Ниже в таблице 1 в качестве примера приводятся правила выбора алломорфов аффиксов множественного числа.

---

[6] Сандхи (санскр. saṅdhí) — изменения звуков на морфемных швах и границе двух слов, объясняемые отчасти фонетически, отчасти как отражение исторических явлений в языке (Лингвистический энциклопедический словарь. Москва, 1990. С. 432).

[7] На вокализм аффиксов в хакасском языке, исходя из закона сингармонизма, влияет главным образом основа слова, а на их консонантные варианты, согласно правилам ассимиляции и аккомодации, помимо основы оказывают влияние также и предшествующие аффиксы. Поэтому помимо сочетаемости основ и аффиксов должна быть проработана фонетическая сочетаемость аффиксов между собой для каждого возможного их комплекта.

Таблица 1

Правила выбора аффиксов множественного числа

№	Фонетические свойства предшествующего форманта	Num(Pl)
1.	а) В элементах, предшествующих аффиксу, последний из охарактеризованных по ряду гласных заднеряден (*а, ы, о, у*) & б) предшествующий элемент оканчивается на гласный или на звонкий неносовой согласный (*г, й, л, р*).	*лар*
2.	а) В элементах, предшествующих аффиксу, последний из охарактеризованных по ряду гласных переднеряден (*е, и, і, ӧ, ӱ*) & б) предшествующий элемент оканчивается на гласный или на звонкий неносовой согласный (*г, й, л, р*).	*лер*
3.	а) В элементах, предшествующих аффиксу, последний из охарактеризованных по ряду 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*
4.	а) В элементах, предшествующих аффиксу, последний из охарактеризованных по ряду гласных переднеряден (*е, и, і, ӧ, ӱ*) & 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*
5.	а) В элементах, предшествующих аффиксу, последний из охарактеризованных по ряду гласных заднеряден (*а, ы, о, у*) & б) предшествующий элемент оканчивается на носовой согласный (*м, н, ң*).	*нар*
6.	а) В элементах, предшествующих аффиксу, последний из охарактеризованных по ряду гласных переднеряден (*е, и, і, ӧ, ӱ*) & б) предшествующий элемент оканчивается на носовой согласный (*м, н, ң*).	*нер*

2.4. Правила сандхи, актуальные для компьютерного морфологического анализа

Правила сандхи подразделяются на два типа: первые описывают морфонологические изменения, возникающие вследствие взаимодействия между морфемами: между основой и аффиксами, между соседними аффиксами; правила второго типа — «поверхностные» правила.

К первому типа относится, например, правило выпадения беглого гласного (*у, ы, і*) при присоединении аффиксов принадлежности, начинающихся на *-ы, -і*: ***пурун/пурн****ы* ‘нос’, ***харын/харн****ы* ‘живот’, ***орын/орн****ы* ‘место’.

Примеры правил второго типа — озвончение глухого согласного в интервокальной позиции — действует всегда (*пас* ‘голова’/*пазым* ‘моя голова’, *абыт-* ‘качать люльку’ (основа)/*абыдарға* (инфинитив)); выпадение согласного в интервокальной позиции и стяжение двух обрамлявших его гласных в один долгий. Чаще всего выпадают звонкие *г, ғ* (практически регулярно). Правила действуют по-разному в зависимости от различных фонетических условий (одно- или многосложная основа, долгота-краткость обрамляющих гласных).

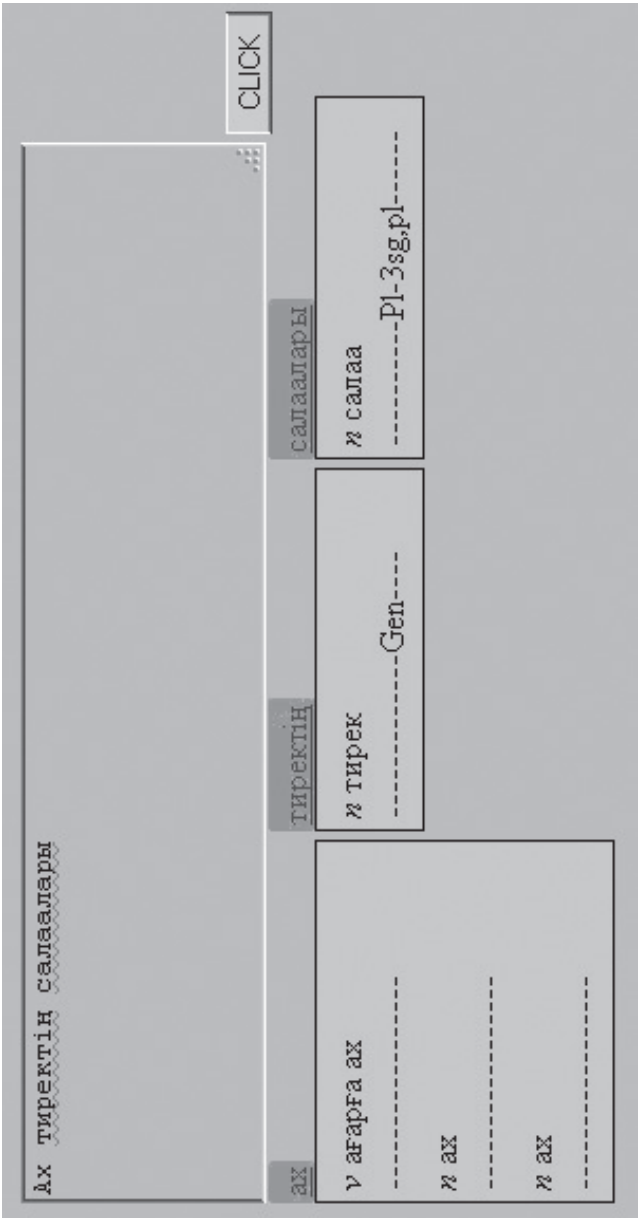
3. Алгоритм морфологического анализа

Анализ словоформы идет справа налево. Сначала программа ищет в словаре основ целую словоформу. Если ее там не оказывается, парсер ищет с правого конца словоформы словоизменительный формант и, если обнаруживается последовательность символов, похожая на какой-либо аффикс из базы, она отрезается, а левая часть снова сравнивается со словарем основ. При отрицательном результате программа снова обращается к правому концу слова и ищет следующий словоизменительный аффикс, сравнивая его с базой аффиксов. Так продолжается до тех пор, пока

оставшаяся слева часть словоформы не совпадет со словом из словаря основ.

Таким образом получается корпус с неснятой грамматической омонимией. Омонимия снимается вручную, а результаты анализа выдаются пользователю в формате html, см. например, анализ строки из эпоса «Ай-Хуучин» (рис. 5).

Относительно возможностей парсера в смысле метрики получения разборов словоформ проблема распадается на две части. Во-первых, вопрос, какому проценту словоформ в тексте парсер приписывает анализ. Этот процент довольно высок (около 95 %), причем отказы парсера обычно обусловлены недостатками не морфологической модели, а словаря (слова нет в словаре, во вхождении словаря неправильно выделена основа, слово диалектное и для него в словаре не выписаны правильные чередования и т.п.). Во-вторых, насколько часто парсер дает словоформе правильный анализ. Обычно среди множества альтернативных анализов правильный имеется. Опять же, обратное случается в основном вследствие несовершенства словаря. Проблема вычленения правильного морфологического анализа пока ложится на постредактирование. Сейчас уже понятно, что ряд неадекватных анализов легко будет устранить при подключении синтаксического модуля; кроме того, планируется подключить семантический модуль в смысле извлечения из семантической разметки словаря (включающей актантные структуры) сведений о семантической сочетаемости внутри предложения.


Ах тиректің салаалары
CLICK
ах
тиректің
салаалары
ү ағарға ах

и ах

и ах

и тирек
----------Gen---
и салаа
----------Pl-3sg,Pl-----

Рис. 5

I. МОДЕЛЬ ИЗМЕНЯЕМОЙ ХАКАССКОЙ СЛОВОИЗМЕНИТЕЛЬНЫХ

№	0	1	2	3	4	5	6	7	8	9
	R (S)	Distr	Form 1	Ptcl 1	Perf, Prosp	Dur	Neg/ Form 2	Iter	Tense (Past, Future), Conv, Mood	Evid
№ п/п										
1.		*КлА*	Form1 *(Б)*	Emph *ТАА*	Perf *(I) бІС*	Dur *чА(Т)*	Neg *БА*	Iter *дІр*	Praes *чА*	Indir *ТІр*
2.			Form. Neg *Бин*	Delim *ЛА*	Prosp (диал.) *(А)К*		Neg.Fut БА.С		Past *КАн*	
3.				Ass *ОК*		Dur1 *и(р)* (для *пар-* и *кил-*)	Neg.Conv *Бин*			
4.							Neg.Conv.Abl *Бин.Аң*		Fut *Ар*	
5.							Form 2 *А*		Hab *ЧАң*	
6.									RPast *ТІ*	
7.									Cunc *КАлАК* Cond *СА*	
8.									Opt *КАй*	
9.										
10.									Conv1 *(І)Б*	
11.									Conv1 *АбАС* диал.	
12.									Conv2 *А*	
13.									Lim *КАли*	
14.										
15.										
16.										
17.										
18.										
19.										

ПРИЛОЖЕНИЯ

СЛОВОФОРМЫ И НАБОР ХАКАССКИХ АФФИКСОВ

Таблица 2

10	11	12	13	14	15		16	17	18	19
Irr	Comit	Num (Pl)	Poss	Apos	Case		Attr	Ptcl 2	Person	Adv
					Simple declension	Possessive declension				
ЧІК	*ЛІГ*	*ЛАр*	1pos.sg *(І)м*	*Ни*	Gen *НІң*	Gen *НІң*	*КІ*	Ass *ОК*	Полные	*Ли*
									1prs.sg *БІн*	
			2pos.sg *(І)ң*		Dat *(К)А*	Dat *(К)А*			2prs.sg *СІң*	
			3pos *(з)І*		Acc *НІ*	Dat *нА*			3prs *дІр*	
			1 pos.pl *(І)бІС*		Loc *ТА*	Acc *Н(І)*			1prs.pl *БІС*	
			2 pos.pl *(І)ңАр*		Abl *ДАң*	Loc *(н)ТА*			2prs.pl *САр*	
					All *САр*	Abl *ДАң*			3prs.pl *СІң.нАр*	
					Instr *нАң*	All *(н)САр*			Краткие	
									1prs.sg *м*	
					Prol *ЧА*	Instr *нАң*			2prs.sg *ң*	
					Delib *ДАңАр*	Prol *(н)ЧА*			2prs.pl *(І)ңАр*	
					Comp *ТАГ*	Delib *ДАңАр*			3prs.pl *ЛАр*	
					Temp *(І)н*	Comp *(н)ДАГ*			Imp.1prs.sg *им*	
						Temp *(І)н*			Imp.3prs.sg *СІн*	
									Imp.1prs.dual *Аң*	
									Imp.1prs.pl *и.бІС*	
									Imp.1prs.pl. Incl *АңАр*	
									Imp.2prs.pl *(І)ңАр*	
									Imp.3prs.pl *СІн.нАр*	
									Prec.1prs.sg *им.дАК*	
									Prec.2prs.sg *ТАК*	

№	0	1	2	3	4	5	6	7	8	9
№ п/п	R (S)	Distr	Form 1	Ptcl 1	Perf, Prosp	Dur	Neg/ Form 2	Iter	Tense (Past, Future), Conv, Mood	Evid
20.										
21.										
22.										
23.										
24.										
25.										

Необходимые пояснения к таблице

По горизонтали в таблице расположен перечень грамматических категорий, выраженных аффиксами, следующими за корнем (основой), по вертикали — аффиксы-показатели грам. категорий со всеми алломорфами.

Одну позицию по горизонтали могут занимать аффиксы одной или нескольких грам. категорий, каждая из которых (кроме R(S)) может быть не выражена, например: единственное число, именительный падеж имени, показатель 3 лица 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 служило критерием невключения того или иного показателя (например, аффиксков деятеля) в разряд словоизменительных.

Distr — дистрибутив, обозначает множественность субъекта или объекта действия

Form (1, 2) — формообразующий аффикс

Form.Neg — отрицательная форма формообразующего аффикса

Ptcl — частица (пишущаяся слитно со словоформой, а также вставная)

10	11	12	13	14	15		16	17	18	19
Irr	Comit	Num (Pl)	Poss	Apos	Case		Attr	Ptcl 2	Person	Adv
					Simple declension	Possessive declension				
									Prec.3prs.sg *СІн.дАК*	
									Prec.1prs.dual *Аң.дАК*	
									Prec.1prs.pl *и.бІс.тАК*	
									Prec.1.prs.pl. Incl *АңАр.дАК*	
									Prec.2prs.pl *(І)ңАр.дАК*	
									Prec.3prs.pl *СІн.нАр.дАК*	

Emph — эмфатическая частица (*и, же, ведь*...)

Delim — делимитативная (ограничительная) частица (*только, лишь*...)

Ass — ассертивная (утвердительная) частица (*также, тоже* …)

Perf — перфектив (завершенность действия)

Prosp — проспектив (состояние, предшествующее действию), исключительно сагайский диалект

Dur — дуратив (длительность действия), также выражает настоящее время

Iter — итератив (периодичность действия) , также выражает настоящее время

Neg — отрицание

Neg.Fut — отрицательная форма буд.вр.

Neg.Conv (1, 2, 3) — отрицательная форма деепричастия

Neg.Conv.Abl — деепричастие мгновенного следования (аблатив отрицательного деепр.)

Tense, Mood — время, наклонение

Fut — будущее время

Hab — хабитуалис (привычное действие в настоящем и прошедшем времени)

Past — прошедшее время

RPast — недавно прошедшее время

Cunc — кункатив, еще не совершившееся действие

Conv (1, 2, 3) — деепричастие

Lim — деепричастие предела в прошлом

Cond — условное наклонение

Assum — ассумптив, предположительное наклонение, вводится оборотом «похоже, что …»

Imp — императив (повелительное наклонение)

Opt — оптатив (желательное наклонение)

Prec — прекатив (просительное наклонение)

Evid — эвиденциальность

Indir — индиректив, косвенная эвиденциальность (неочевидность либо заглазность) действия

Irr — «ирреалис», сослагательное наклонение

Comit — комитатив, показатель совместности («с …», «вместе с …»)

Num (Sg, Pl, Dual) — число (ед., мн., двойств.)

Poss, pos — принадлежность

Apos — показатель посессивного имени. После этого аффикса имя (или субстантивированное причастие) 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)

Список падежей

Nom — номинатив, или нулевой падеж, отсутствует в таблице, т.к. не имеет поверхностного выражения

Gen — генетив (родительный)

Dat — датив

Acc — аккузатив (винительный)

Loc — локатив (местный)

Abl — аблатив (исходный)

All — аллатив (направительный)

Instr — инструментальный (творительный)

Prol — пролатив (продольный)

Delib — делибератив (выраж. объект речи или мысли, также причинно-следственные отношения)

Comp — компаратив (сравнительный)

Temp — темпоральный (падеж временного обстоятельства)

Attr — атрибутивный показатель

Adv — адвербиальный показатель

Person, prs — лицо

Attr и Adv — показатели, которые функционируют обычно как словообразовательные, прибавляясь к чистой основе; тогда соответствующая единица попадает в словарь, и аффиксы не обязательно учитываются при морф. анализе. Но оба они могут свободно присоединяться к концу длинной словоформы, содержащей уже словоизменительные аффиксы, при помещении ее в определенные синтаксические условия; тогда их нет в составе словарной словоформы. В этом случае парсер подвергает эти аффиксы анализу.

Кумулятивно выраженные граммемы разделяются точками.

Выделены специальные морфонемы на месте фонем, чередующихся в зависимости от фонетических позиций. Аффиксы в таблице записаны с помощью морфонем, что позволяет не выписывать в каждом случае весь набор алломорфов того или иного аффикса:

Согласные морфонемы	Гласные морфонемы
Б: б/п/м	*А: е, а*
К: *г/г/х/к*	*І: і, ы*
Г: г/г/0	*О: о, ӧ*
Т: т/д	
Д:т/д/н	
С: с/з	
Л: л/н/т	
Н: н/т	
Ч:ч/ҷ	
Ң: ң/0	

Гласный в скобках, стоящий в начале морфа, проясняется, если предыдущий морф кончается на согласный, и опускается, если предыдущий морф кончается на гласный. Согласный в скобках, стоящий в начале морфа, проясняется, если предыдущий морф кончается на гласный, и опускается, если предыдущий морф кончается на согласный.

Литература

Большой хакасско-русский словарь. Под ред. *О.В. Субраковой*. Новосибирск, 2006.

Бречалова Е.В. Принципы построения синтаксического представления корейского предложения. Дисс. на соискание ученой степени кандидата филологических наук. Москва, 2009.

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

Грамматика хакасского языка. Под ред. *Н.А. Баскакова*. Москва, 1975.

Жолковский А.К., Мельчук И.А. О семантическом синтезе // Проблемы кибернетики. Вып. 19. Москва, 1967.

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

Крылов С.А. Использование системы StarLing при создании морфологически аннотированного корпуса современного монгольского языка. 2011. *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. Москва, 2005. С. 111–135.

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

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

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

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

Хакасско-русский словарь. Под ред. *Н.А. Баскакова*, с приложением грамматического очерка хакасского языка *Н.А. Баскакова*. Москва, 1953.

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

Martin G. A Reference Grammar of Korean. Chicago, 1992.

Анна Владимировна Дыбо
Институт языкознания РАН
Москва, Россия
Anna Vladimirovna Dybo
Institute of Linguistics at the Russian Academy of Sciences
Moscow, Russia
adybo@mail.ru

Александра Валерьевна Шеймович
Институт языкознания РАН
Москва, Россия
Alexandra Valerievna Sheimovich
Institute of Linguistics at the Russian Academy of Sciences
Moscow, Russia
asheimovich@yandex.ru