V МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ «TURKLANG 2017»

Труды конференции

Том 2

КАЗАНЬ
2017

УДК 004.8+81’32
ББК 81.1

Организаторы:

Академия наук Республики Татарстан
Институт прикладной семиотики

Казанский (Приволжский) федеральный университет
Высшая школа информационных технологий
и информационных систем
Институт вычислительной математики
и информационных технологий

Евразийский национальный университет имени Л. Н. Гумилёва Министерства образования и науки Республики Казахстан
НИИ «Искусственный интеллект»

Международная Тюркская академия

Российская ассоциация искусственного интеллекта

Издание осуществлено при финансовой поддержке
Российского фонда фундаментальных исследований
(проект №17-47-161033)

Научные редакторы:
академик АН РТ, профессор, д.т.н. Д. Ш. Сулейманов,
к.т.н. А. Р. Гатиатуллин

Пятая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2017». – Труды конференции. В 2-х томах. Т 2. – Казань: Издательство Академии наук Республики Татарстан, 2017. – 327 с.
ISBN 978-5-9690-0407-8

Сборник содержит материалы Пятой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2017» (Казань, Татарстан, Россия, 18–21 октября 2017 г.)

Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной лингвистики и ее приложений.

УДК 004.8+81’32
ББК 81.1

ISBN 978-5-9690-0407-8

УДК 81'33

EXPERIENCE OF COMPUTER-ORIENTED DESCRIPTION OF THE TUVAN MORPHOPHONOLOGY WITHIN THE FRAMEWORK OF THE SYSTEM OF AUTOMATIC MORPHOLOGICAL ANALYSIS

A. Dybo, A. Sheymovich
Institute of linguistics of RAS, Moscow, Russia
adybo@mail.ru, asheimovich@yandex.ru

This paper describes brief results of the most recent stage of the work on the development of an automatic morphological analyzer for Tuvan language, which was started in a frame of the Presidium of RAS project “Corpus Linguistics”, and resumed in line with the DHPS RAS program “The working out, creation and developing of electronic parallel linguistic corpora of minority Turkic languages and dialects of Russia”. The primary purpose of the present paper is the ordering of segment rules of the Tuvan wordform generation as well as their interpretation in the parser design.

Keywords: wordform generation, segment rules of allomorph choise, morphological rules, morphophonological rules, phonological rules, graphic rules, automatic morphological analyser.

ОПЫТ КОМПЬЮТЕРНО-ОРИЕНТИРОВАННОГО ОПИСАНИЯ ТУВИНСКОЙ МОРФОНОЛОГИИ В РАМКАХ СИСТЕМЫ АВТОМАТИЧЕСКОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА[1]

А.В. Дыбо, А.В. Шеймович
Институт языкознания РАН, Москва
adybo@mail.ru, asheimovich@yandex.ru

В работе представлены краткие результаты очередного этапа работы над корпусными технологиями, начатыми в рамках программы Президиума РАН «Корпусная лингвистика» и продолжающимися по проекту программы ОИФН РАН «Разработка, создание и развитие электронных параллельных лингвистических корпусов миноритарных тюркских языков и диалектов РФ». В настоящей статье основное внимание уделено классификации сегментных правил синтеза словоформы тувинского языка и тому, каким образом эти правила интерпретируются в теле разрабатываемого парсера.

[1] Работа выполнялась по гранту РГНФ № 15-04-00370 «Разработка анкет для сбора материалов к интегральному описанию миноритарных тюркских языков и диалектов России».

Ключевые слова: синтез словоформы, сегментные правила выбора алломорфов, морфологические правила, морфонологические правила, фонологические правила, графические правила, автоматический морфологический анализ.

Введение. О наших проектах

Уже несколько лет развиваются исследования по программам РГНФ (РФФИ), Президиума РАН и ОИФН РАН «Корпусная лингвистика», направление «Корпуса языков народов России». В рамках проекта по корпусам миноритарных тюркских языков создан и пополняется новыми материалами корпус хакасского языка, насчитывающий к настоящему времени около 500 тыс. словоформ. В 2014 г. на научно-практическом семинаре UniTurk "Унификация систем грамматической разметки в корпусах тюркских языков" в рамках конференции TEL2014 в Казани была представлена рабочая модель автоматического парсера для корпуса хакасского языка, опирающаяся на компьютерную модель хакасской словоформы (Дыбо, Шеймович 2014); также продолжается грамматическое, словообразовательное и семантическое аннотирование хакасского корпуса. Наработки в области автоматической разметки хакаского корпуса мы планируем использовать при создании парсеров для тувинского (в настоящее время) и шорского (в ближайшем будущем) языков.

Примерно с того же времени (2012 г.) коллектив научного центра при Тувинском госуниверситете в Кызыле работал над созданием тувинского корпуса и предпринимал шаги по направлению к его лингвистическому аннотированию. Разработки А.Б.Хертек и Б.Ч. Ооржак в области создания модели тувинской словоформы были представлены на конференции TurkLang-2015 (Oorzhak, Khertek 2015). С использованием этих разработок на основе инструментария Helsinki Finite-State Toolkit (HFST) американскими тюркологами из университета Блумингтона был построен автоматический анализатор для тувинского языка (Washington, Bayyr-ool 2016) Наша модель в целом близка предложенной американскими коллегами, но отличается в нескольких существенных моментах:

1. Наш анализатор обрабатывает словоформу не слева направо, а справа налево, что по нашему мнению, лучше соответствует реальному устройству тюркской морфологии.

2. В анализаторе используется грамматический тувинско-русский словарь, существующий в виде электронной базы данных с грамматической, морфонологической, словообразовательной и семантической разметками, сформированный на основе ТРС 1968.

3. В отличие от Washington, Bayyr-ool 2016, мы не используем двух отдельных алгоритмов для анализа глагольных и именных словоформ. Используется единый алгоритм, не выстраивающий парадигмы. Каждая грамматическая категория считается либо выраженной, либо не выраженной, т.о. мы избегаем нулевых аффиксов. См. ниже.

4. Пока не анализируются композитные слова, пишущиеся через пробел, и аналитические словоформы, в т.ч. личные формы глагола. Проблема определения аналитических личных форм глагола, так же как и проблема любых аналитических форм должна решаться на этапе синтаксического анализа с использованием информации об относительном порядке синтетических словоформ.

1. Грамматика

По аналогии с компьютерно-ориентированной моделью хакасской словоформы, с опорой на принципы грамматики порядков мы разработали модель тувинской словоформы и собрали инвентарь словоизменительных показателей для тувинского языка (Дыбо, Шеймович 2015). Для этого использован традиционный инвентарь словоизменительных морфем из Грамматики тувинского языка (ГТЯ 1961) и очерка Ш.Ч. Сата к Тувинско-русскому словарю (Сат 1955). Эта работа была сделана с учетом исследований коллег из научного центра при Тувинском госуниверситете (Хертек, Ооржак 2012; Oorzhak, Khertek 2015; Salchak, Bayir-ool 2015). В настоящей статье представлена несколько доработанная версия модели тувинской словоформы (см. Приложения, табл. 3) с упором на описание сегментных преобразований, с которыми приходится сталкиваться парсингу, оказавшееся куда более сложным, чем для хакасского: процессы ассимиляции согласных (например, при выпадении беглого гласного) последовательно отражаются в тувинской орфографии, что представляет определенные проблемы для автоматического морфологического анализатора.

Имея опыт построения порядковой модели хакасской словоформы, можно сказать о главных отличиях ее от тувинской.

Порядковая модель тувинской словоформы устроена проще хакасской в основном благодаря не так далеко зашедшей фузии внутри сложной глагольной словоформы. Во-первых, в тувинском лицо и число при главной предикации в значительной части случаев (при именах и причастиях) выражаются аналитически, с помощью местоименных частиц («личных показателей» по Сат 1955, 685), по форме совпадающих с личными местоимениями. Внутрь синтетической словоформы попадает только кумулятивное выражение 3 лица ед.числа в форме презенса трех глаголов: *тур-* ‘стоять’, *олур* – ‘сидеть’, *чор* – ‘ходить’; кумулятивное выражение лица и числа при претерите на *-ды* и кумулятивное выражение лица, числа и наклонения в аффиксах императива, условного и предельного наклонений. Во-вторых, в хакасском в синтетическую глагольную словоформу включился ряд акциональных аффиксов, восходящих к вспомогательным глаголам (как дуратив на *-чат-* или презенс на *-ча*, восходящие к глаголу **jat-* ‘лежать’), которые в тувинском сохранили свой аналитический статус. Ср.: хак. *ойнапчабыс* ‘мы играем’ vs тув. *ойнап тур бис* ‘мы играем’.

На первый взгляд тувинская модель проще, в ней меньше мест, пока не предвидится дублирования глагольных слотов.

2. Словарь

Для работы нашего парсера необходим словарь в форме электронной базы данных, содержащей частеречные пометы и указания на чередования основ, не описанные фонологическими правилами, а также некоторую другую информацию (семантическую, словообразовательную, этимологическую и пр.). К настоящему времени средствами СУБД Starling конвертирован в словарную базу данных Тувинско-русский словарь под ред. Э.Р.Тенишева (ТРС 1968), включающий более 17 тыс. лексем. Образцом для словарной статьи тув. словарной базы послужила уже апробированная форма статьи базы данных на основе БХРС. Частей речи в ней выделяется столько же, сколько для хакасского языка: Имя (Nomen), Глагол (Verbum), неизменяемое (Invar). О выделении грамматических классов и частей речи в тюркской морфологии применительно к нуждам компьютерной обработки текста неоднократно говорилось в процессе разработки хакасского парсера и

модели тувинской словоформы, поэтому здесь мы ограничимся отсылками к (Шеймович 2012а-б; Дыбо, Шеймович 2014; Дыбо, Шеймович 2015).

3. Разграничение словообразовательной и словоизменительной морфологии

Морфологическая разметка содержит информацию о словоизменительных, но не о словообразовательных признаках лексемы. Деривативы, записанные в словаре, не являются предметом грамматического парсинга (в словарной базе данных имеются соответствующие поля, содержащие расчлененную морфонологическую запись (DERIV) и словообразовательную разметку (DERIVGLOSS) основы). Специфический вопрос, который здесь встает, это трактовка залогов. Вообще говоря, залог в грамматиках тюркских языков часто принято описывать как словоизменительную категорию. Это связано, во-первых, с относительной регулярностью способов его образования, во-вторых, с синтаксическим характером его грамматического значения. Причины, по которым мы оставляем залог за пределами грамматического анализа, относя его к словообразованию, были неоднократно указаны в предшествующих работах, посвященных хакасскому парсеру (Шеймович 2012а-б; Дыбо, Шеймович 2014 и след.). В тюркологии давно показано, что невозможно выделить порядковое место залога в словоформе (Циммер 1987). Нередки случаи, когда наличие формального показателя залога в слове не связано с наличием у него соответствующего значения: например, основы, содержащие словообразовательный каузативный аффикс, не всегда являются каузативами семантически[1].

[1] См. ГТЯ 276: «Глаголы в понудительном залоге, образованные от переходных глаголов, могут быть как переходными, так и непереходными. В первом случае они имеют понудительное значение, во втором – страдательное. Переходность (понудительное значение) или непереходность (страдательное значение) таких глаголов зависит от контекста. Примеры: 1) Ол алгыны меңээ эттеткен ‘Ту шкуру дали выделать мне’ (понудительное значение); Меңээ эттеткен алгы бо-дур ‘Вот это выделанная мной шкура’ (страдательное значение); 2) Хойларны бөрүге өлүртпес ‘Не давать волкам резать овец’ (понудительное значение); Бөрүге өлүрткен хоювус ийи ‘Волком зарезано две наших овцы’».

Помимо залоговых аффиксов в сфере словообразования нами оставлены (вслед за ГТЯ и в соответствии с наличием этих образований в ТРС и ТСТЯ), например, аффикс «желательности» *(Ы) КсА* (ГТЯ 269) (*алыксаар /ал=ЫКсА-/* желат. от *ал-* (см. *алыр*) ‘хотеть брать’: *бо номну алыксап тур мен* ‘я хочу взять эту книгу’; *суксаар /су=КсА-/* ‘жаждать, сильно хотеть пить’) и аффикс «прекратительности» *БАстА* (ГТЯ 409) (*тоовастаар /тоо=БАстА-/* ‘переставать обращать внимание’, прекрат. от *тоор /тоо-/* ‘обращать внимание’: *Бистиң чамдык аңчыларывыс дииң болгаш ас, күзен тоовастаан* ‘У нас некоторые наши охотники белку и горностая, колонка перестали ценить’ (ТСТЯ 251)).

4. Сегментные преобразования в тувинской грамматике

Что касается сегментных преобразований, то обработка явлений сингармонизма уже была формализована для хакасского парсера. Эти наработки с незначительными изменениями (такими, как наличие губных вариантов (отсутствующих в хакасском) для словоизменительных аффиксов с узким гласным) использованы для тувинского анализатора.

Основной сложностью, с которой сталкивается построение тувинского парсера, в отличие от хакасского, является описание тувинских сандхи – морфонологических процессов, происходящих на границах морфем, т.е. изменений, возникающих вследствие взаимодействия между контактными морфемами: между основой и аффиксами, между соседними аффиксами – с целью последующей их переработки в правила для автоматического глоссирования тувинских текстов.

Как уже было сказано выше, обработка сандхи в тувинском парсере гораздо сложнее, чем в хакасском, в частности, поскольку процессы ассимиляции согласных последовательно отражены в тувинской орфографии (в отличие от хакасской).

Ниже перечислены основные правила сегментных преобразований, действующие в тувинском. Мы излагаем правила поуровнево, в том порядке, в котором они следуют в грамматике синтеза словоформы, поскольку анализ осуществляется «через синтез»,

т.е. путем многократных прямых и обратных проходов с проверкой гипотез[1].

0. **Сегментные правила выбора алломорфов (поверхностно-морфологические правила выбора одного из алломорфов морфемы, при заданном морфемном составе словоформы)**:

0.1. Гласная в скобках, стоящая в начале морфа, обозначает, что если предыдущий морф кончается на согласную, будет выбран вариант данного морфа, начинающийся на эту гласную морфонему; а если предыдущий морф кончается на гласную, будет выбран вариант данного морфа, начинающийся на согласную морфонему (т.е. гласная в скобках опускается). Согласная в скобках, стоящая в начале морфа, обозначает, что если предыдущий морф кончается на гласную, будет выбран вариант данного морфа, начинающийся на эту согласную морфонему; а если предыдущий морф кончается на согласную, будет выбран вариант данного морфа, начинающийся на гласную морфонему (т.е. согласная в скобках опускается).

0.2. Алломорфы, стоящие в одной клетке таблицы 3 через запятую (Fut *ЫР, Ар*; ConvPraes *А, ЫЛ*), выбираются согласно информации в грамматическом словаре основ, в поле ALTERNATEN.

0.3. Алломорф ConvPraes *-й* выбирается в случае, если предыдущий морф оканчивается на гласную (*бода-* ‘думать’ – *бода-й* ‘думая’, *тары-* ‘сеять, сажать’ – *тары-й* ‘сея, сажая’).

1. **Правила морфологических чередований**, применение которых требует информации из словарей морфем (из грамматического словаря основ – например, сведения о лексико-грамматическом классе основы или словарные данные этой основы, – и

[1] Для облегчения использования человеком электронного грамматического словаря тувинского языка мы приняли решение выписывать в морфонологической транскрипции, т.е. с использованием условных букв, только звуковую оболочку аффиксов, а не корней (морфонологическая транскрипция словообразовательных аффиксов дается внутри расчлененной морфонологической записи основы в поле словарной базы DERIV). Это решение является чисто формальным. Отсутствие условной морфонологической записи корней компенсируется информацией из полей словарной базы ALTERNAT, ALTERNATEN и FORM; поскольку соответствующая информация является релевантной для небольшого числа основ, такое решение представляется оправданным. Ниже в тексте статьи мы приводим примеры основ в условной морфонологической записи.

из таблицы словоизменительных аффиксов – например, граммема данной морфемы).

1.1. Преобразование конечных (стоящих непосредственно перед морфологической словоизменительной границей[1]) согласных *-к, -К* неодносложной последовательности морфем на границе с аффиксом принадлежности (*Ы, Ым*, *Ың, Ывыс, ЫңАр*) в *Г*: *тавак + Ым > таваГ-Ым* ‘мое блюдо’; *белек + Ым > белеГ-Ым* ‘мой подарок’, *кезек*[2] *+ Ы > кезеГ-Ы* (*> кезээ*) ‘его часть’ (для односложной последовательности правило не работает: *аак + Ы > аак-Ы* ‘его последствие’). Информация о наличии чередования извлекается из сегментного состава морфем, стоящих перед морфологической границей, и сегментного состава и имени аффикса принадлежности.

1.2. Преобразование конечных (стоящих непосредственно перед морфологической словоизменительной границей) *к, п, м, л*[3] некоторых односложных глагольных основ на границе с афф. деепричастия на *Ып* в *Г*: *бол- + Ып > боГ-Ып/бол-Ып (> бооп/болуп)* ‘быв’, *кел- + Ып > кээп* ‘придя’, но *бил- + Ып > билип* ‘узнав’. Информация о наличии и обязательности преобразования извлекается из полей ALTERNATEN или FORM базы грамматического словаря основ и сегментного состава и имени аффикса деепричастия[4].

---

[1] В условной морфонологической расчлененной записи граница между словообразовательными и словоизменительными показателями различается (у нас это «=» для словообразовательных и «-» для словоизменительных). Формулируемое правило учитывает только «-».

[2] В условной записи в поле DERIV *кеС=Ак*, но на вход при порождении поступает лексическая основа.

[3] По данным ГТЯ 38, опционально это правило действует также для односложных глаголов на *-ң* (*доң-Ып > дооп/доңуп* ‘замерзнув’), но по примерам в ТСТЯ все такие основы сохраняют *ң*. Вопрос нуждается в корпусном и диалектологическом исследовании.

[4] На этом этапе технически имеет смысл просто извлекать готовую форму деепричастия из поля словарной базы. Дело в том, что один из глаголов на *-л* (*ал-* ‘брать’), все глаголы на *-к*, для которых релевантно выпадение, и по крайней мере часть глаголов на *-п* получают после выпадения согласного в форме деепричастия на *-Ып* краткую, а не долгую гласную: *ап* ‘взяв', *хып* ‘загоревшись’ (*хывар*), *соп* ‘ударив’ (*согар*). К сожалению, полная информация по этим формам отсутствует и в грамматиках, и в словарях тувинского языка; в части случаев извлекается из примеров в словарных статьях ТСТЯ. Вопрос нуждается в корпусном исследовании.

1.3. Преобразование *р, л* некоторых односложных глагольных основ на границе с афф. деепричастия будущего времени *Ыр*, *Ар* в *Г*: *кел- + Ар > кеГ-Ар/кел-Ар* (*> кээр/келир*) ‘приходить’, *бер- + Ар > беГ-Ар* (*> бээр*) ‘давать’. Информация о наличии и обязательности чередования извлекается из поля ALTERNATEN базы грамматического словаря основ и имени аффикса деепричастия.

1.4. «Беглые» гласные.

Узкая гласная *Ы* закрытого конечного слога последовательности морфов, если этой гласной предшествует одиночный согласный (т.е. в последовательности *(С)VCЫC*), если эту последовательность не разрывает словоизменительная морфологическая граница, может выпадать, если к ней присоединяется аффикс, начинающийся на гласную. Эти процессы зависят от словарной информации о составляющих форму морфемах, а не только от их поверхностной структуры и морфемных границ, и потому не могут считаться морфонологическими. Ср. разное развитие в однотипных морфонемных последовательностях: *бүрЫн-Ы > бүрүн-үбиле* ‘полностью’ – 3Poss от *бүрүн* ‘всё’ – без выпадения и *бурЫн-Ы > мурну* ‘перед, раньше’ – 3Poss от *бурун* ‘прежний’ (о начальной согласной см. ниже), – с выпадением; *дайын* ‘война’ – *Ада-чурттуң Улуг дайын-ы* ‘Великая Отечественная война’ – без выпадения, и *оюн* ‘игра’ – 3Poss *ойн-у* – с выпадением; инфинитивы *ажын-ар* /*аШы=н-* 1/ ‘сердиться, дуться, злиться’ без выпадения – *ашт-ыр* /*аШы=н-* 2/ ‘оправдаться, доказать свою невинность’ с выпадением (и ассимиляцией, см. ниже). Вид основы с выпавшей гласной указан в статье словарной базы в поле FORM.

Примеры.

Имя:

Выпадение беглой гласной из 2-го закрытого слога двусложной именной основы при присоединении афф. принадлежности: *ойЫн-Ы > ойн-Ы > ойну* ‘его игра’ (при формулировании правила для автоматического парсера графема *ю* раскрывается в виде *йу*, см. ниже); *эрин-Ы > эрн-Ы > эрни* ‘его губа’; *оГЫл-Ы > огл-Ы > оглу* ‘его сын’.

Глагол:

Выпадение беглой гласной из залогового аффикса (иногда уже окаменевшего): *хайын-Ыр > хайн-Ыр > хайныр* ‘кипеть’, *дир=Ыл-Ы > дирл-Ы > дирли* ‘ожив’, *ажы=н=Ыш-Ыр > ажыныш-Ыр > ажынчыр* ‘сердиться друг на друга’.

1.5. Уникальные чередования: основа *бөрт* ‘шапка’ в позиции перед вокалическим началом следующего морфа имеет форму *бөрг-*: *бөрт + Ы > бөрг-ү* ‘его шапка’. Основа *аас* ‘рот’ в той же позиции имеет форму *акс-: аас + Ы > аксы* ‘его рот’[1]. Сведения – в словарной базе, в поле FORM.

2. **Морфонологические правила**, применение которых определяется сегментным составом сочетающихся морфем, включающим информацию о морфемных границах.

2.1. Ассимиляции в сочетаниях согласных морфонем, не разделенных словоизменительной границей. Сочетания согласных морфонем, не разделенные словоизменительной морфологической границей, претерпевают следующие ассимиляции:

2.1.1. (Если при присоединении аффиксов выпадает узкая гласная, стоявшая в позиции между согласными *л-н,* то) сочетание согласных *лн > нн*:

Имя: *келин + Ы > келн-Ы > кенн-Ы > кенни* ‘его невестка’;

Глагол: *лVн > нн*: *кыл=ын- + Ыр > кылн-Ыр > кынн-Ыр > кынныр* ‘сделаться’;

2.1.2. (Если при присоединении аффиксов выпадает узкий гласный, стоявший в позиции между согласными $С_{сильн}Vн$, $С_{сильн}$-*л*, то) *CVн, CVл > Ст*[2]:

Имя: *иШин + Ы > иШн-Ы > ишт-Ы > ишти* ‘его живот’, *эКин + Ы > эКн-Ы > экт-Ы > экти* ‘его плечо’;

Глагол: *саКын-* ‘вспоминать’ + *Ып> сактып*; *оТун-* 'просыпаться' + *Ыр > оттур, тыП=ыл-* ‘находиться’ + *Ыр > тыптыр*

2.1.3. (Если при присоединении аффиксов выпадает узкий гласный, стоявший в позиции между согласными $С_{сильн}Vш$, или

---

[1] Эта основа должна была бы получить условную запись аГЫс, но по обычным правилам выпадения и ассимиляции мы бы получили *агз-ы, ср. процессы в глаголе дагзыр /даас-/ ‘обязывать кого-либо’.

[2] Таким образом, очевидно, что действие правила выпадения узкой гласной в тувинском предшествует действию правила интервокального озвончения.

лVш, то) *CVш > Сч: диК=иш-* ‘помочь строить’ + *ЫР >дикчир*[1], *кыП=ыс+ЫI > кыпсы* ‘зажигая’. (Если при присоединении аффиксов выпадает узкий гласный, стоявший в позиции между согласными *рVш,* то) *рVш > рж: таПар=ыш-* ‘встречаться’ + *ЫР > таваржыр.*

С сандхи типа 2.1. связан еще один тип фонетических изменений, дистантная регрессивная ассимиляция согласной в начале слова, вызванная последней согласной в пределах одного закрытого слога:

2.2.1. В именных и глагольных основах вышеописанного типа с начальной *б-*: *бил=ин* ‘признаваться’ + *ЫР > билн-ЫР > биннЫР* (контактная регрессивная ассимиляция по назальности) *> миннир*; *бурЫн-ЫI > мурну* ‘перед, раньше’ – 3 Poss от *бурун* ‘прежний’. Это чередование также отражается в словаре в поле ALTERNATEN.

2.2.2. Чередование согласных в начале односложных глагольных основ, оканчивающихся на сильный согласный и с фарингализованным гласным при присоединении к ним аффиксов на гласную: *хыъп* ‘гори!’ + Fut *Ар > кыъвар* ‘гореть’, *төък* 'вылей!' + Fut *Ар > дөъгер* ‘выливать’. Это чередование лишь частично отражено в тувинской орфографии (см. Пальмбах 1956, 111-112), а именно отражено для заднеязычных и не отражено для зубных согласных. Оно также сводится к дистантной ассимиляции начального согласного согласному, закрывающему слог, по силе/глухости/придыхательности в условиях наличия фарингализованной гласной. Мы записываем его в поле ALTERNATEN в случае отражения в орфографии.

2.3. Ассимиляция согласных, разделенных словоизменительной морфемной границей (см. табл. 1).

Аффиксы, начинающиеся с определенной морфонемы, получают реализацию этой морфонемы в зависимости от качества конеч-

---

[1] В словообразовательной морфонологии (т.е. в соседстве с морфемными границами типа =) сочетаемость согласных морфонем отличается, и, в частности, имеются пары слов с разным поведением одного и того же аффикса, привязанным к разнице значений (*садыг=жы* ‘торговец’, ‘купец’ // *садык=чы* ‘продавец’; ср. *садыг* ‘торговля’, ‘продажа’, *сат-* ‘торговать’, ‘продавать’; *тараа=жы* ‘хлебороб’ // *тараа=чы* ‘тот, кто едет за хлебом’; ср. *тараа* ‘хлеб’), см. ГТЯ 149. Частично это объясняется бытованием монголизмов на *-чи*, см. ГТЯ 169.

ной морфонемы предшествующего морфа[1] (инвентарь словоизменительных морфем в морфонологической условной записи см. в таблице 3; условная запись прочих морфем извлекается из полей лексической базы ALTERNAT, ALTERNATEN, или DERIV; либо, если эти поля в словарной статье не заполнены, то совпадает с начальной формой слова).

Аффиксы с начальной морфонемой *Б-* получают *в* после гласной, *б* после слабой неносовой согласной, *п* после сильной, *м* после носовой;

Аффиксы с начальной морфонемой *К-* получают *г* после слабой согласной или гласной, *к* после сильной согласной.

Аффиксы с начальной морфонемой *Г-* получают 0 после гласной (выпадение с последующим стяжением гласных вокруг морфемной границы в долгую *АА,* **с последующим сингармоническим преобразованием в зависимости от рядности гласной основы)**: *хову + ГА > ховаа* ‘к степи’, *кижи + ГА > кижээ* ‘человеку’*; сана- + ГАн > санаан* ‘сосчитал’, *номчу- + ГАн > номчаан* ‘прочитал’, *бижи- + ГАн > бижээн* ‘написал’, *чүлү-+ ГАш > чүлээш* ‘побрив’, *сөгле- + ГАй > сөглээй* ‘можешь сказать’, *ойна- + ГАлА > ойнаала* ‘как только поиграл’*, номчу- + ГЫже > номчааже* ‘пока не прочитает’); *г* после слабой согласной, *к* после сильной согласной.

Аффиксы с начальной морфонемой *Т-* получают *д* после слабой согласной или гласной (*уруг-да* ‘в ребенке’), *т* после сильной согласной (*тавак-та* ‘в блюде’).

Аффиксы с начальной морфонемой *Н-* получают *д* после слабой неносовой согласной или гласной, *н* после носовой согласной, *т* после сильной согласной.

Аффиксы с начальной морфонемой *С-* получают *з* после слабой согласной или гласной, *с* после сильной согласной.

Аффиксы с начальной морфонемой *Л-* получают *л* после гласной или слабой неносовой согласной (но не *л*), *т* после сильной, *н* после носовой; *д* – после *л*.

Аффиксы с начальной морфонемой *Ч-* получают *ч* после сильной согласной, а также после сонорных согласных *л, м, н, ң*; *ж* – после гласных и слабой *г,* а также после сонорных *й, р.*

---

[1] Эти процессы связаны с наличием фонологических запретов на сочетаемость согласных, см. табл. 4 в приложении.

Таблица 1

Начальные морфонемы афф. Конечные морфонемы основы	**Б**	**К**	**Т**	**Н**	**С**	**Л**	**Ч**
V	в	г	д	н	з	л	ж
т, Т	п	к	т	т	с	т	ч
п, П	п	к	т	т	с	т	ч
м	м	г	д	н	з	н	ч
н	м	г	д	н	з	н	ч
ң	м	г	д	н	з	н	ч
л	б	г	д	д	з	д	ч
р	б	г	д	н	з	л	ж
й	б	г	д	д	з	л	ж
с, С	п	к	т	т	с	т	ч
к, К	п	к	т	т	с	т	ч
г, Г	б	г	д	д	з	л	ж
ч, Ч	п	к	т	т	с	т	ч
ш	п	к	т	т	с	т	ч

3. **Фонологические правила**, включающие исключительно фонологические условия (не зависящие от морфологического членения словоформы). Это фактически правила перехода с морфонемного на фонемный уровень записи, при котором осуществляется пересчет «морфонем» предыдущего уровня в «фонемы». Для парсера, работающего с текстом в орфографической форме, это пересчет с записи, содержащей условные буквы, в просто буквенную запись.

3.0. Все границы морфем ликвидируются. Если с обеих сторон от морфемной границы стояли гласные (уже без скобок – скобки были убраны на этапе 0), то эти гласные стягиваются в одну долгую, **являющуюся соотносительной парой по долготе гласной, предшествовавшей** морфемной границе. На этом этапе, например, происходит стяжение конечного гласного глагольной осно-

вы с первым гласным афф. деепричастия будущего времени *ЫР*, *Ар*: *ойна- + Ар > ойнаар* ‘играть’, *сөгле- + Ар >сөглээр* ‘сказать’, *чыры- + ЫР > чырыыр* ‘светить’ (в словарной базе такие стяженные глагольные формы представлены в поле FIELD1, т.к. деепричастие будущего времени используется в тувинском языке в качестве неопределенной формы глагола (инфинитива)).

3.1. Все *Г, г*, имеющиеся в условной записи словоформы до сих пор и оказавшиеся в интервокальной позиции, выпадают, а окружающие их гласные стягиваются в одну долгую, **являющуюся соотносительной парой по долготе гласной, предшествовавшей выпавшей *Г*:** *оГЫл > оол, аГЫс > аас, таваГ-ЫМ > таваам* ‘мое блюдо’*; уруг-Ы > уруу ‘его ребенок’; белеГ-ЫМ > белээм* ‘мой подарок’; *кес=еГ-Ы > кезээ* ‘его часть’; *даг-Ы > даа* ‘его гора’; *суг-Ы > суу* ‘его вода’; *саг-ЫП > саап* ‘подоив’, *каг-Ар > каар* ‘оставлять’, *чуг-ЫР > чуур* ‘мыть’, *кеГ-Ар/кел-Ар > кээр/ келир* ‘приходить’*, беГ-Ар > бээр* ‘давать’.

3.2. Ослабление (графическое озвончение) глухой согласной при попадании ее в интервокальную позицию. Для парсинга существенно это озвончение на границе основы и аффикса, начинающегося на гласную: *ат* ‘имя’ – *ад-ы* ‘его имя’, *час* ‘весна’ – *чазын* ‘весной’, *аак-Ы* ‘его последствие’ > *аагы*: *соок аагы* ‘последствия мороза’; но в принципе это правило действует независимо от прохождения морфемных границ: *эКЫН* ‘плечо’ > *эгин, иШЫН* ‘живот’ > *ижин*.

3.3. Рядный и губной сингармонизм (см. подробное описание ГТЯ 46-49). Большинство словоизменительных аффиксов имеют заднерядный, и переднерядный, огубленный и неогубленный варианты, которые выбираются в соответствии с характеристикой последней гласной словоизменительной основы по рядности и огубленности. Аффиксы, содержащие гласную морфонему А, выбирают алломорф с *а* после заднерядной гласной; с *е* после переднерядной гласной. Аффиксы, содержащие гласную морфонему Ы, выбирают алломорф *ы* после заднерядной неогубленной гласной; *и* после переднерядной неогубленной гласной, *у* после заднерядной огубленной гласной, *ү* после переднерядной неогубленной гласной. Это правило в принципе действует независимо от границ и состава морфем в словоформе, так что его можно переформулировать так:

Таблица 2

$\mathbf{V_n}$ / $\mathbf{V_{n-1}}$	**А**	**Ы**
V_{back}	а	ы
V_{front}	е	и
$V_{front\ lab}$	е	ү
$V_{back\ lab}$	а	у

3.4. Все сохранившиеся на этом этапе условные буквы (здесь в морфонологической транскрипции применены как условные кириллические буквы верхнего регистра) преобразуются в соответствующие буквы нижнего регистра, таким образом возвращая орфографическую форму тувинской словоформы (до обработки графическими правилами).

4. **«Графические сандхи»**, обусловленные способом употребления кириллической графики (запись сочетаний звуков вида *й+V* с помощью кириллических йотированных букв), например:

4.1. *йу > ю, йа > я, йе > е* (ГТЯ 43).

Примеры:

хой ‘овца’ – Poss.1Pl *хоювус* ‘наша овца’ < *хой-ЫвЫс*;

ой- ‘прорубить, пробивать’, Prosp *ойгаш* < *ой-КАш*, инфинитив *ояр*< *ой-Ар*, ConvPast *оюп* < *ой-Ып* (ТСТЯ 481);

өй- ‘валять, катать’, Prosp *өйгеш* < *өй-КАш* и инфинитив *өер* < *өй-Ар,* ConvPast *өйүп* < *өй-Ып* (ТСТЯ 490).

4.2. Сочетание букв *ее*, возникшее при порождении словоформы, на поверхностном уровне заменяется на ээ: *кел+Ыр > кеер >* Fut *кээр* ‘приходить’ (см. правило 1.3).

Заключение

Можно видеть, что сегментные процессы, обусловливающие синхронный облик тувинской словоформы, представляют довольно сложную, многоуровневую систему, которая может быть проинтерпретирована как конгломерат разновременных фонетических процессов, постепенно вытеснявшихся на более глубинные уровни языка, а также вымывавшихся аналогическими процессами. Ср. рассуждения о «частичном восстановлении исторического

облика» некоторых тувинских слов в определенных морфологических условиях в ГТЯ 118. Автоматический морфологический анализатор может обойти часть этих явлений с помощью «лексических исключений», поместив соответствующую информацию в словарные статьи лексической базы, но значительное их число все-таки должно обрабатываться алгоритмом. Отметим, что существующие словари и грамматики тувинского языка содержат значительные лакуны в описании сегментного поведения ряда морфем и типов морфем, которые могут быть закрыты только с помощью корпусного исследования.

ЛИТЕРАТУРА

1. Aelita Salchak, Aziyana Bayir-ool. The main results of the project on creation an electronic corpus of Tuvan language // Сборник трудов конференции «TurkLang-2015». Казань, 2015. С. 259–268.

2. Oorzhak B., Khertek A. Development of semantyc markup the corpus of Tuvan language // Сборник трудов конференции «TurkLang-2015». Казань, 2015. С. 351–362.

3. Gleason 1955 – Gleason, H. Introduction to descriptive linguistics. New York, 1955: Holt, Rinehart and Winston.

4. ГТЯ – Исхаков Ф.Г., Пальмбах А.А. Грамматика тувинского языка. М.,1961. 473 с.

5. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Филология и культура. 2014. № 2 (36). С. 20–26;

6. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов хакасского и древнетюркского языков // Научное обозрение Саяно-Алтая. Серия: Филология. № 8, 2014. С. 9–30.

7. Дыбо А.В., Шеймович А.В, Крылов С.А. Некоторые возможности семантической и этимологической разметки для корпусов тюркских языков (расстановка семантических тэгов в электронном хакасско-русском словаре // Сборник трудов международной конференции TurkLang 2015. Казань, 2015. С. 304–327. 488 с.

8. Дыбо А.В., Шеймович А.В. Порядковая модель тувинской словоформы // Материалы региональной конференции «Языки народов Сибири и сопредельных регионов». Новосибирск, ИФ СО РАН, 6–9 октября 2015 г.

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

10. Пальмбах 1956 – Пальмбах А.А. Система согласных тувинского языка и ее отражение в письменности. // Ученые записки Тув. научно-исслед. института языка, литературы и истории. IV. Кызыл, 1956.

11. ТРС 1968 – Тувинско-русский словарь / Под ред. Э.Р. Тенишева. М., 1968.

12. Сат 1955 – Сат Ш.Ч. Тувинский язык (краткий очерк) // Тувинско-русский словарь / под ред. А.А. Пальмбаха. М., 1955. С. 613–721.

13. ТСТЯ – Толковый словарь тувинского языка. Т. II. Новосибирск 2011.

14. Хертек А.Б., Ооржак Б.Ч. О морфологической разметке электронного корпуса текстов тувинского языка // Филологические науки. Вопросы теории и практики. Тамбов. 2012 г. № 7 (18). Ч. II. С. 214–218.

15. Циммер К. О некоторых ограничениях на каузативизацию в турецком языке // Новое в зарубежной лингвистике. Т. XIX. 1987. С. 283–298.

16. Шамина Л.А. Аналитические грамматические формы и конструкции в функции сказуемого в тувинском языке. Новосибирск, 2010 г., 240 с.

17. Шеймович 2012а – Шеймович А.В. Некоторые особенности автоматического анализа морфологии хакасского языка (на материале корпуса) // д. на Международном научном форуме «Н.Ф.Катанов и современность». Абакан, 2012.

18. Шеймович А.В. О принципах построения автоматического морфологического анализатора для корпуса хакасского языка // д. на Отчетном собрании Российского комитета тюркологов. Москва, 2013.

19. Washington, Bayyr-ool 2016 – Washington J. N, Bayyr-ool A. et al. The development of a finite-state morphological analyser for Tuvan // Родной язык, № 1(4), М., 2016.

Таблица 3. **Модель изменяемой тувинской глагольной**

0	1	2	3	4	5
S	**Distr**	**Perf**	**Neg**	**Tense (Praes, Past, Future) Mood**	**Num(Pl)**
	Distr *КЫлА*	Perf *(Ы)ББЫт*	Neg *БА*	Praes *ыр*	Pl *ЛАр*
			Neg.Conv *БАйн*	Past *ТЫ*	
			Neg.Fut *БАс*	PastRes *ГАн*	
				PastRtr *ЧЫк*	
				Cont *ББЫшААн*	
				Fut *Ыр, Ар*	
				Cunc *ГАлАк*	
				ConvLim *ГАлА*	
				ConvPast *(Ы)п*	
				ConvPraes *А, Ы, й*	
				Prosp *ГАш*	
				ConvLim2 *ГЫже*	
				Abstr *КЫ*	
				Opt. *ГАй*	

ПРИЛОЖЕНИЯ

словоформы и набор тувинских словоизменительных аффиксов

6	7		8	9
Poss	**Case**		**Person (1, 2)**	**Ptcl**
	Simple declencion	**Possessive declencion**		
1pos.sg *(Ы)м*	Gen *НЫң*	Gen *НЫң*	1sg *м*	Interr *(Ы)л*
2pos.sg *(Ы)ң*	Dat *ГА*	Dat *нГА*	2sg *ң*	Add *-даа*
3pos. *(з)Ы*	Acc *НЫ*	Acc *Н*	1pl *ББІс*	Indir *-ТЫр*
1pl *(Ы)ББІс*	Loc *ТА*	Loc *нТА*	2pl *ңАр*	Emph *-Ла*
2 pl *(Ы)ңАр*	Abl *ТАн*	Abl *нТАн*	3pl *ЛАр*	
Gen.3Pos *НЫЫ*	Lat1 *Че*	Lat1 *нЧе*	Praes.3sg *у*	
	Lat2 *ТЫвА*	Lat2 *нТЫвА*	Imp.1sg *ЫЙн, Айн*	
	Lat3 *КЫдЫ*	Lat *КЫдЫ*	Imp.3sg *СЫн*	
			Imp.1.Dual *ААл(Ы), ЫЫІл(Ы)*	
			Imp.1.pl *ААлЫңАр, ЫЫІлЫңАр,*	
			Imp.2.pl *(Ы)ңАр*	
			Imp.3.pl *СЫн.(нАр)*	
			1.sg.Lim *м.че*	
			2.sg.Lim *ң.че*	
			1.pl.Lim *вис.че*	
			2.pl.Lim *ңер.же*	
			1.sg.Cond *СЫ.м.зА*	
			2.sg.Cond *СЫ.ң.зА*	
			3sg,pl Cond *зА*	
			1.pl.Cond *СЫ.вЫс.сА*	
			2.pl.Cond *СЫ.ңАр.зА*	

Условные обозначения

Abstr – абстрактное имя действия (герундий)

Add – аддитивная частица (*и, ни, же, ведь…*)

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)

Lat – латив (направительный)

Instr – инструментальный (творительный)

Cond – условное наклонение

Conv – деепричастие

ConvLim – деепричастие предела в прошлом

Cunc – кункатив, еще не совершившееся действие

Distr – дистрибутив, обозначает множественность субъекта или объекта действия;

Emph – эмфатическая частица (*-mo*)

Fut – будущее время

Imp – императив (повелительное наклонение)

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

Interr – вопросительность

Lim – предельность («Предельное наклонение»)

Mood – наклонение

Neg – отрицание

Neg.Conv – отрицательная форма деепричастия

Neg.Fut – отрицательная форма буд.вр.

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

Opt – желательное наклонение

Past – прошедшее время

Perf – перфектив (завершенность действия)
Person, prs – лицо
Poss, pos – принадлежность
Praes = Pres – настоящее время
Prosp – проспектив (состояние, предшествующее действию)
Ptcl – частица (пишущаяся слитно со словоформой, а также вставная)

S – основа. Основа включает корень со словообразовательными показателями и присутствует 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.

Tense – время

Кумулятивно выраженные граммемы разделяются точками.

Значение условных символов в системе морфонем:

Согласные морфонемы
Б: б/п/м/в
К: *г/к*
Г: *г/к/0*
Т: т/д
Н:д/т/н
С: с/з
Л: л/т/д/н
Ч:ч/ж/ш

Гласные морфонемы
А: е, а
Ы: и, ы, у, ү

Таблица 4. **Фрагмент системы допустимых сочетаний согласных в тувинском языке**

	т	д	л	н	с	з	ч	ш	ж
с	кес-тик	–	–	–	бас-сын	–	кас-чыр-	–	–
т	эът-тиг	–	–	–	четсе	–	бутчак	–	–
к	өөк-тээр	–	–	–	сукса-	–	көрүк-чү	акша	–
ш	чеш-тин-	–	–	–	ыш-сыг	–	дашчы	–	–
р	эрте	көрдү	баар-лыг	силер-ниң	–	барза	кадар-чы	–	хааржак

л	–	мал-дың	–	–	–	эл-зиит-	болчур	–	–
м	–	амдан вкус	–	эмне-	–	хүлүм-зүр-	кымчы	–	–
н	–	ындыг	–	хүннер	–	хүнзе-	кинчи	–	–
ң	–	аңдар-	–	деңне-	–	даңзы	аңчы	–	–
й	ойта-яр	шай-дан	чай-лаг	ойнаар	–	дай-зын	–	–	хоорай-же, далай-жы
г	–	дагда	суг-лук	дагны	–	аарыг-зыыр	садыг-чы	–	дагже

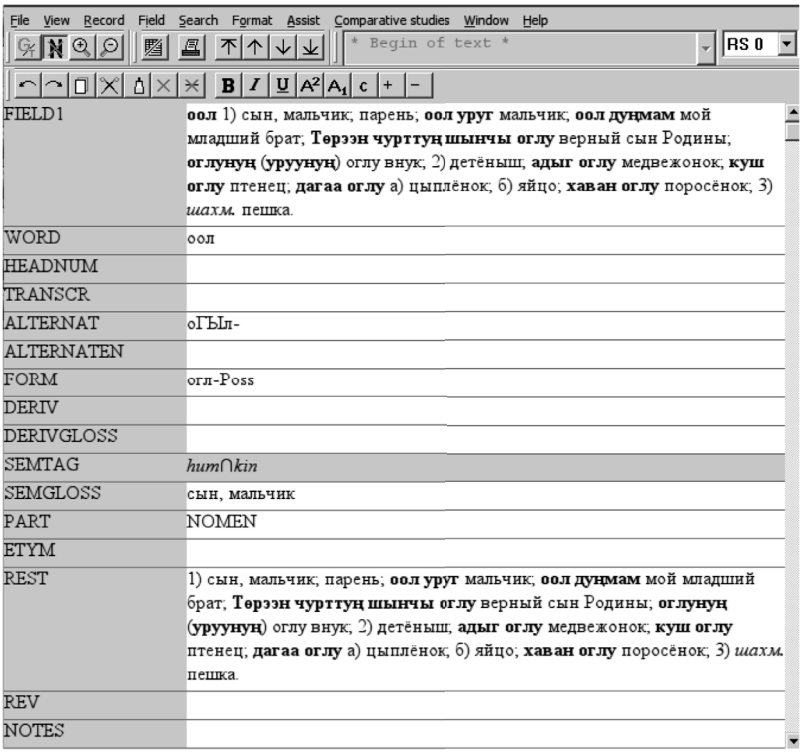

Рис. 1. Образец именной статьи тувинско-русской электронной базы данных

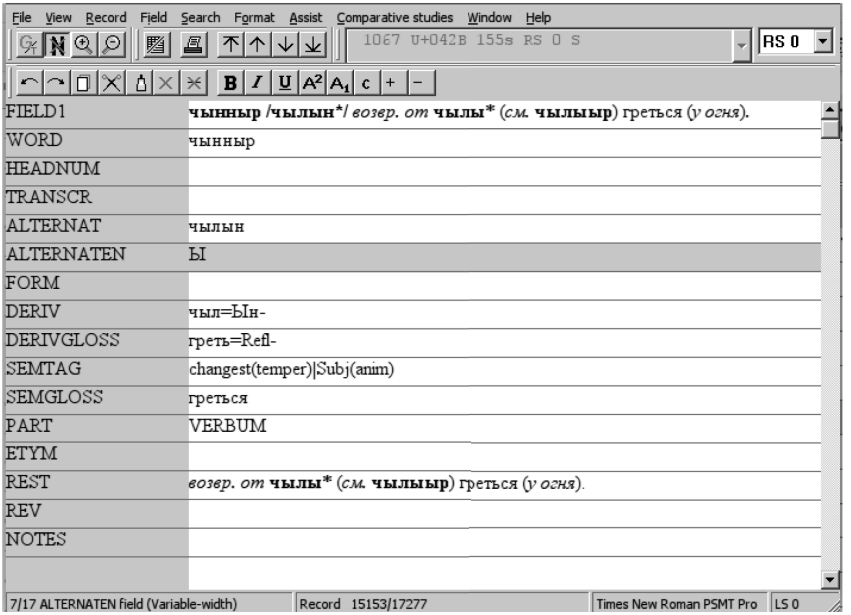

Рис. 2. Образец глагольной статьи тувинско-русской электронной базы данных