УДК 519.7
DOI: 10.15827/2311-6749.17.2.7

ЭЛЕКТРОННЫЙ КОРПУСНЫЙ СЛОВАРЬ ТУВИНСКОГО ЯЗЫКА

А.С. Дагбажык, аспирант, angyrool-d@mail.ru;
Ч.М. Монгуш, аспирант, mongushchod91@yandex.ru
(Сибирский федеральный университет, просп. Свободный, 79, г. Красноярск, 660041, Россия)

Словари – важная часть всякого корпуса естественного языка. Под корпусом понимается информационно-справочная система, основанная на собрании оцифрованных текстов. Словари в корпусах, как правило, многофункциональны и состоят из словарных статей. В работе предложена модель представления словарной статьи в БД. Показана версия программных средств ведения корпусного словаря тувинского языка в Microsoft Office Access. Разработанные программные и информационные средства предназначены для изучения тувинского языка с точки зрения написания, произношения и толкования, а также для организации поиска слов и словосочетаний в текстах, хранящихся в корпусе.

Ключевые слова: *организация словарей, корпусы естественно-языковых текстов.*

В настоящее время активно создаются корпусы естественных языков с помощью современных информационных технологий и методов математического моделирования. Под корпусом понимается информационно-справочная система, основанная на собрании оцифрованных текстов. Корпус включает в себя различные письменные и устные тексты, представленные в данном языке, различные типы словарей, а также разметку – информацию о свойствах текстов. Разметка отличает корпус от электронных библиотек текстов. В корпусах используются следующие типы разметки: метатекстовая, морфологическая, синтаксическая, семантическая и др. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. С помощью корпусов решаются многие филологические и лингвистические задачи.

Для многих языков народов Российской Федерации, в том числе для тюркских языков, создаются национальные корпусы. Работа над формированием Национального корпуса тувинского языка ведется преподавателями, аспирантами и студентами Тувинского государственного и Сибирского федерального университетов. На сегодняшний день в Национальном корпусе тувинского языка содержатся тексты тувинской художественной литературы (прозы, поэзии, драматургии, фольклора), официально-деловых документов. В корпус также входят частотный словарь по художественным произведениям на тувинском языке, тувинско-русский электронный словарь «ТывЛин», словарь диалектных слов алтайского диалекта тувинского языка, составленные М.В. Бавуу-Сюрюн и С.М. Далаа [1, 2].

Словари – важная часть корпуса. Различают несколько типов словарей. Словари в корпусах (или корпусные словари) наделены многими функциями. Корпусный словарь содержит всю лингвистическую информацию о каждом слове. Различают следующие типы словарей: переводные, грамматические, орфографические, орфоэпические, словообразовательные.

Переводные словари содержат сопоставление слова одного языка с его переводом на другом языке. Переводные словари условно разделяют на две большие группы: общелексические словари, которые переводят общую лексику с одного языка на другой язык; научные и научно-технические словари, которые охватывают специальные термины по определенным отраслям знаний. Грамматические словари содержат сведения о морфологических и синтаксических свойствах слова. Морфология изучает части речи, их категории и формы слов, а синтаксис – структуру словосочетаний и предложений. Орфографические словари определяют правила написания слов. Орфоэпические словари указывают правила произношения слов. Словообразовательные словари отражают словообразовательную структуру слов. В них слова приводятся с ударением и разделяются на морфемы. Под морфемой принято понимать минимальную значимую часть слова. К морфемам относятся корни (или основы) и аффиксы. Корень – морфема, несущая лексическое значение слова или основную часть этого значения. Аффикс – морфема, видоизменяющая значение слова или основную часть этого значения. К аффиксам относятся префиксы и постфиксы [3, 4]. Корпусный словарь сочетает в себе, как правило, функции переводного, грамматического, орфографического, орфоэпического и словообразовательного типов словарей.

В организационной структуре различных типов словарей много общего. Основу всякого словаря составляют словарные статьи. Каждая словарная статья посвящена отдельному слову, его называют заголовочным (или заглавным). Словарная статья характеризует заголовочное слово с соответствующих точек зрения. Например, для англо-русских переводных словарей Л.П. Ступиным определены следующие части словарной статьи: *entry word* – заголовочное слово, *sense* – смысл, *definition* – определение переводного эквивалента, *quotation* – цитирование, *reference* – отсылка, *status label* – метка о временной или территориальной ограниченности употребления слова, *regional label* – метка о территориальной употре-

бительности слова, *functional label* – метка о принадлежности слова к части речи, *subject label* – метка о принадлежности слова к определенной области знаний [5].

Корпусный словарь тувинского языка разрабатывается для изучения тувинского языка с точки зрения написания, произношения, толкования, а также для организации поиска слов и словосочетаний в текстах, хранящихся в корпусе. Поэтому корпусный словарь тувинского языка сочетает в себе функции переводного, грамматического, орфографического, орфоэпического типов словарей. Для организации данного словаря выбрана структура словарной статьи, позволяющая хранить соответствующую лексическую информацию о заголовочном слове (рис. 1).

Заголовочное слово
Перевод на русский, английский, немецкий языки
Транскрипция
Основа слова
Лемма (нормальная форма слова)
Граммема (грамматическая характеристика) или морфологическое описание слова (часть речи, число и другие характеристики)
Парадигмы – все формы слова. Каждой форме сопоставлены грамматические признаки, в частности, число, падеж, лицо, время, наклонение
Значение слова
Этимологическая справка (сведения о происхождении слова)
Метка о принадлежности к аббревиатурам
Метка о наличии синонима, омонима и антонима
Дополнительная информация

Рис. 1. Структура словарной статьи

При разработке корпусного словаря тувинского языка учитывались особенности морфологической структуры этого языка. С точки зрения морфологических особенностей различают изолирующие языки (морфемы максимально отделены друг от друга), агглютинативные (морфемы семантически отделены, но реально объединены в слова), флективные (семантические и формальные границы между морфемами плохо различимы) [6].

В изолирующих языках слово обычно совпадает с корнем, а отношения между словами задаются порядком слов, служебными словами, ритмом, интонацией. Например, к изолирующим языкам относится китайский язык. В агглютинативных языках доминирующим способом словоизменения является агглютинация (от лат. *agglutinatio* – склеивание). Структура слова в агглютинативных языках характеризуется большим количеством аффиксов, прибавляемых к неизменяемой основе слова. Каждый из аффиксов имеет только одно, строго назначенное значение. Располагаются аффиксы в определенном порядке, но вместе с тем некоторые из них могут быть пропущены. К агглютинативным языкам относятся турецкий и казахский языки. Для флективных языков характерны многозначность морфем и сильная связь морфем в слове, не позволяющая им сравнительно свободно передвигаться внутри слова, как в агглютинативных языках. Типичными примерами флективных языков являются немецкий и русский. Тувинский язык относится к ветви тюркских языков, а с точки зрения морфологии к агглютинативным [3]. Морфология тувинского языка соответствует следующим требованиям: фиксированная последовательность словообразовательных аффиксов, их грамматическая однозначность, однократность появления в данной словоформе аффикса определенной граммемы, низкий процент грамматической омонимии, отсутствие префиксов [7, 8].

Основные сведения в каждой отдельной словарной статье корпусного словаря тувинского языка составляют морфологические признаки заголовочного слова. Эти сведения извлекаются из распознанных и выправленных копий тувинско-русских и других переводных, морфологических и орфографических словарей тувинского языка, а также имеющихся БД основ и множества обнаруженных аффиксов тувинского языка [1, 2, 7]. Предполагается, что пополнение корпусного словаря осуществляется в процессе стемматизации (нахождения основы) новых заголовочных слов, их лемматизации (приведения к нормальной или словарной форме), выявления граммем и получения парадигм. Разработка методов и средств реализации этих процессов составляет один из отдельных этапов совершенствования Национального корпуса тувинского языка, выполняемого при поддержке Российского гуманитарного научного фонда (грант 16-34-1-01033).

Приведем описание БД и разработанную версию программных средств создания и ведения корпусного словаря тувинского языка средствами Microsoft Office Access. БД корпусного словаря включает сле-

дующие реляционные таблицы: MAIN – основная таблица с заголовочным словом; RUS, ENG, GER – таблицы с переводом заголовочного слова на различные языки (русский, английский, немецкий); MORPHOLOGY – таблица, содержащая морфологические признаки заголовочного слова. Структура этих таблиц представлена на рисунках 2–5.

main

Имя поля	Тип данных	Описание (необ
entry_id	Счетчик	идентификатор для заголовочного слова
article	Короткий текст	заголовочное слово
mean	Длинный текст	значение слова
transcription	Короткий текст	транскрипция слова
example	Длинный текст	примеры применения слова в предложениях и в речи
speech	Числовой	помета о принадлежности части речи

Рис. 2. Структура таблицы MAIN

rus

Имя поля	Тип данных	Опи
rus_id	Счетчик	идентификатор для слова на русском
russian	Короткий текст	слово на русском языке
entry_id	Числовой	используется для связи с таблицей main

Рис. 3. Структура таблицы RUS

eng

Имя поля	Тип данных	Опис
eng_id	Счетчик	идентификатор для слова на английском
english	Короткий текст	слово на английском языке
entry_id	Числовой	используется для связи с таблицей main

Рис. 4. Структура таблицы ENG

morphology

Имя поля	Тип данных	Опи
entry_id	Числовой	используется для связи с таблицей main
case_1	Короткий текст	падеж 1
case_2	Короткий текст	падеж 2
case_3	Короткий текст	падеж 3
case_4	Короткий текст	падеж 4
case_5	Короткий текст	падеж 5
case_6	Короткий текст	падеж 6
case_7	Короткий текст	падеж 7
case_8	Короткий текст	падеж 8
case_9	Короткий текст	падеж 9

Рис. 5. Структура таблицы MORPHOLOGY

Работа с корпусным словарем тувинского языка осуществляется с помощью следующих функций: добавление и редактирование статьи, удаление статьи, поиск словарной статьи с транскрипцией, формирование и визуализация морфологических признаков заглавного слова. Соответствующие интерфейсы этих функций приведены на рисунках 6–9.

Структура словарной статьи и ее представление в БД Microsoft Office Access позволяют при необходимости расширять функции корпусного словаря и использовать этот словарь

– для изучения тувинского языка с точки зрения написания, произношения, толкования, перевода на другие языки;

– для формирования морфологической, синтаксической и семантической разметки тувинских текстов;

– для организации поиска слов и словосочетаний в текстах, хранящихся в Национальном корпусе тувинского языка.

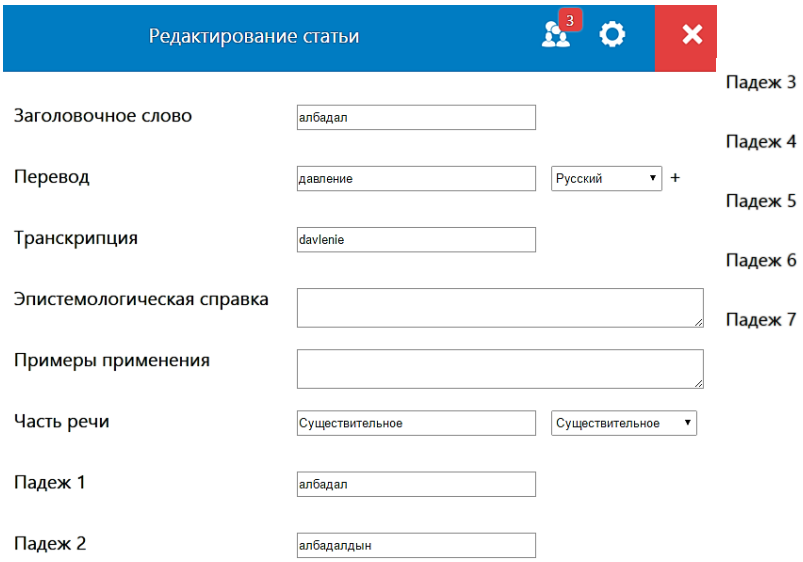

Рис. 6. Добавление и редактирование словарной статьи

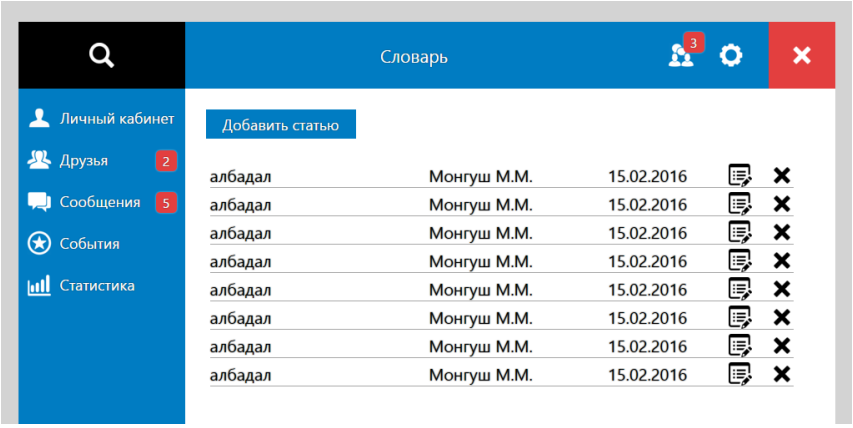

Рис. 7. Удаление словарной статьи

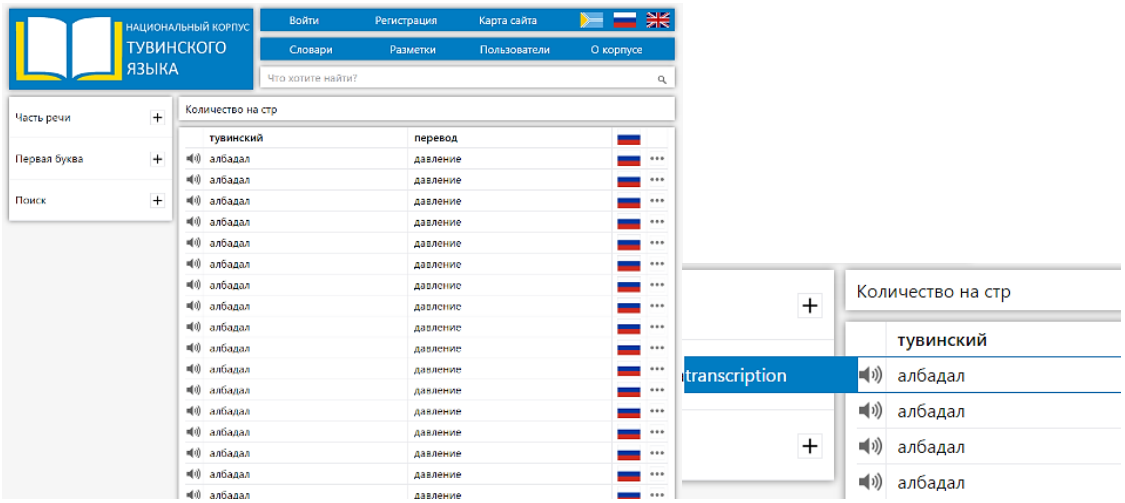

Рис. 8. Поиск словарной статьи с транскрипцией

Перевод		
аргументация	барымдаазын көргүзер	барымдаалаар

Грамматические признаки

Род		
Мужской (masculinum)	Женский (femininum)	Средний (neutrum)
Число		
Единственное (Singular)	Множественное (Plural)	Собирательное (Multiplex)
Лицо		
Первое	Второе	Третье
Падеж		
Именительный		
Родительный		
Дательный		
Винительный		
Творительный		
Предложный		

Морфологические признаки

Часть речи
Имя существительное
Имя прилагательное
Имя числительное
Наречие
Местоимение
Глагол
Причастие
Деепричастие

Рис. 9. Морфологические признаки заголовочного слова

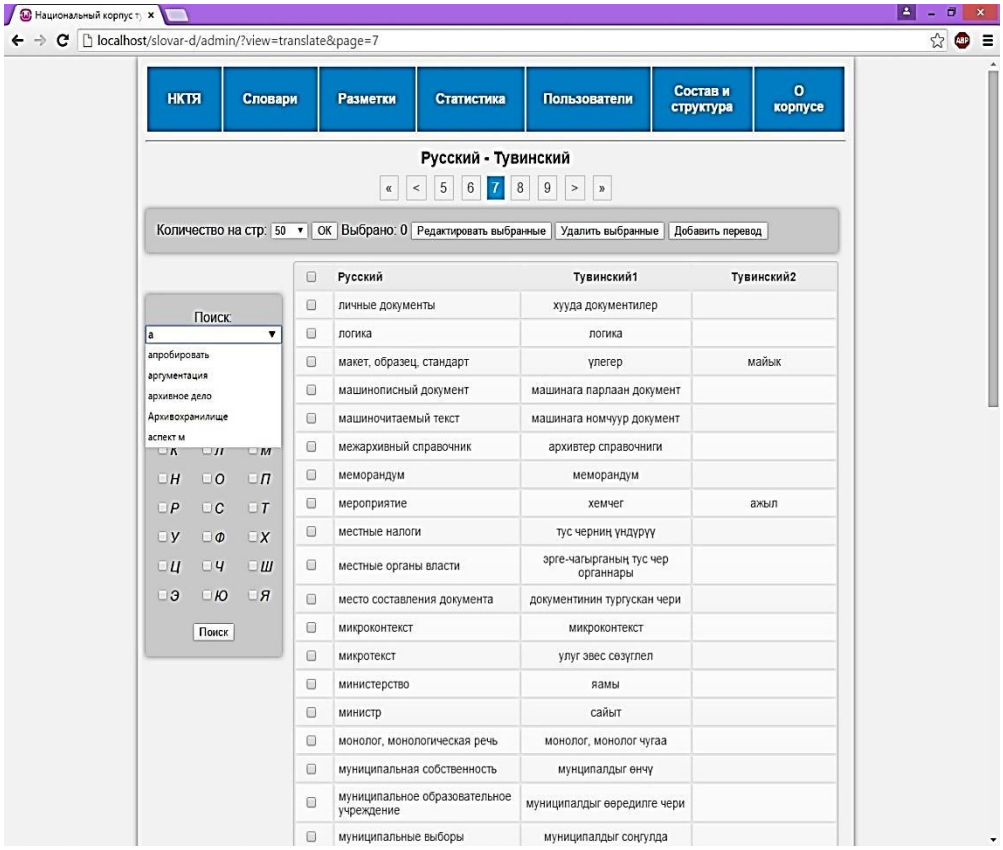

Рис. 10. Интерфейс менеджера корпусного словаря

В настоящее время разрабатывается менеджер корпусного словаря, реализующий различные функции по извлечению из словаря необходимой информации: поиск конкретного слова, поиск слова по лемме, поиск слов по набору морфологических признаков и другим позициям словарной статьи. Интерфейс ме-

неджера корпусного словаря представлен на рисунке 10. Также ведутся работы по заполнению корпусного словаря, созданию программных средств стемматизации и лемматизации заголовочных слов, извлекаемых из текстов тувинской художественной литературы (прозы, поэзии, драматургии, фольклора) и официально-деловых документов на тувинском языке.

Литература

1. Бавуу-Сюрюн М.В. Тувинский язык на современном этапе. URL: http://www.tuva.asia/journal/issue_7/2158-bavyy-suyruyn-mv.html (дата обращения: 10.04.2017).
2. Салчак А.Я., Байыр-Оол А.В. Электронный корпус тувинского языка: состояние, проблемы // Мир науки, культуры, образование. 2013. № 6. С. 408–409.
3. Исхаков Ф.Г., Пальмбах А.А. Грамматика тувинского языка. Фонетика и морфология. М.: Изд-во восточ. лит-ры, 1961. 472 с.
4. Плунгян В.А. Общая морфология: Введение в проблематику. М.: Эдиториал УРСС, 2000. 384 с.
5. Ссорина М.С. Словарь как мультиструктурная организация // Ярославский пед. вестн.: Гуманитарные науки. 2011. № 1. Т. 1. С. 142–146.
6. Батура Т.В. Математическая лингвистика и автоматическая обработка текстов. Новосибирск: Изд-во РИЦ НГУ, 2016. 166 с.
7. Хертек А.Б., Ооржак Б.Ч. О морфологической разметке электронного корпуса текстов тувинского языка. URL: http://www.gramota.net/materials/2/2012/7-2/57.html (дата обращения: 10.04.2017).
8. Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков // Философия и культура. 2014. № 2. С. 20–26.