СТРУКТУРА СЛОВАРНОЙ СТАТЬИ НАЦИОНАЛЬНОГО КОРПУСА ТУВИНСКОГО ЯЗЫКА

THE STRUCTURE OF AN ENTRY IN THE NATIONAL CORPUS OF TUVAN LANGUAGE

Ангыр-оол С. Дагбажык
Сибирский федеральный университет

Angyr-ool S. Dagbazhyk
Siberian Federal University

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, представленных в данном языке, различные типы словарей, а также разметку — информацию о свойствах текстов. Разметка отличает корпус от электронных библиотек текстов.

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.

В статье представлена структура словарной статьи для Национального корпуса тувинского языка. База данных корпусного словаря включает следующие таблицы: MAIN — основная таблица с заголовочным словом; RUS, ENG, GER — таблицы с переводом заго-

Contemporary information technologies and mathematical modelling has made creating corpora of natural languages significantly easier. A corpus is an information and reference system based on a collection of digitally pro-cessed texts. A corpus includes various written and oral texts in the given lan-guage, a set of dictionaries and markup – information on the properties of the text. It is the presence of the markup which distinguishes a corpus from an electronic library.

At the moment, national corpora are being set up for many languages of the Russian Federation, including those of the Turkic peoples. Faculty members, postgraduate and undergraduate students at Tuvan State University and Siberian Federal University are working on the National corpus of Tuvan language.

This article describes the structure of a dictionary entry in the National corpus of Tuvan language.

Дагбажык Ангыр-оол Сымчаан-оолович — аспирант Института математики и фундаментальной информатики Сибирского федерального университета (г. Красноярск). Адрес: 660041, Россия, г. Красноярск, пр. Свободный, 79, ауд. 34-03. Тел.: +7 (391) 206-21-48. Эл. адрес: angyrool-d@mail.ru Научный руководитель — д-р физ.-мат. н., проф. В. В. Быкова.

Dagbazhyk Angyr-ool Symchaan-oolovich, Postgraduate student, Institute of Mathematics and Fundamental Information Technology, Siberian Federal University. Postal address: Room 34-03, 79 Svobodny Pr., 660041 Krasnoyarsk, Russian Federation. Tel.: +7 (391) 206-21-48. Email: angy-rool-d@mail.ru. Research advisor: Doctor of Physics and Mathematics, Pro-fessor V.V. Bykova.

ловочного слова на различные языки; MORPHOLOGY — таблица с морфологическими данными. База данных реализована в Microsoft Office Access.

Для работы с корпусным словарем реализованы следующие функции: добавление новой статьи, редактирование статьи, удаление статьи, поиск словарной статьи с транскрипцией, формирование и визуализация морфологических признаков заглавного слова.

Представленный проект позволяет рассматривать корпусный словарь как мультиструктурную организацию со сложным иерархическим строением, важнейшим корневым компонентом которого является словарная статья. Разработанный корпусный словарь может быть использован 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; корпус текстов; словарь; тувинский язык; электронный словарь; Microsoft Office Access

The corpus database comprises the following tables: MAIN – the headword table, RUS, ENG, GER — translations of the headword into three languages, MORPHOLOGY — the table containing morphological data on the headword. The database is built in Microsoft Of-fice Access.

Working with the corpus dictionary includes the following functions: adding, editing and removing an entry, entry search (with transcription), set-ting and visualizing morphological features of a headword.

The project allows us to view the corpus dictionary as a multi-structure entity with a complex hierarchical structure and a dictionary entry as its key component. The corpus dictionary we developed can be used for studying Tu-van language in its pronunciation, orthography and word analysis, as well as for searching for words and collocations in the texts included into the corpus.

Keywords: dictionary structure; textual corpus; dictionary; Tuvan lan-guage; electronic dictionary; Microsoft Office Access

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, представленных в данном языке, различные типы словарей, а также разметку — информацию о свойствах текстов. Разметка отличает корпус от электронных библиотек текстов (Салчак, Байыр-оол, 2013). Традиционно в корпусах используются следующие типы разметки: метатекстовая, морфологическая, синтаксическая, семантическая и др. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. На основании 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 (Бавуу-Сюрюн, 2010: Электр. ресурс; Салчак, Байыр-оол, 2013).

Словари — важная часть корпуса. Различают несколько типов словарей. Словари в корпусах, как правило, многофункциональны. Корпусный словарь содержит всю лингвистическую информацию о каждом слове. Различают следующие типы корпусных словарей: диалектные, грамматические, орфографические, словообразовательные, переводные.

Диалектные словари — разновидность толковых словарей, описывающих лексику одного или группы говоров (диалектов). Диалект — разновидность данного языка, употребляемая в качестве средства общения между лицами, связанными тесной территориальной, социальной или профессиональной общностью. Грамматические словари — это словари, которые содержат сведения о морфологических и синтаксических свойствах слова. Морфология — раздел грамматики, изучающий части речи, их категории и формы слов. Синтаксис — раздел лингвистики, изучающий строение словосочетаний и предложений. Орфографические словари содержат правила написания слова при письме. Орфоэпические словари — словари, отражающие правила произношения. Орфоэпия — раздел фонетики, занимающийся нормами произношения, их обоснованием и установлением. Словообразовательные словари — словари, отражающие словообразовательную структуру слов. Слова в словообразовательных словарях приводятся с расчленением на морфемы и с ударением.

Переводные словари — словари, содержащие сопоставление слова одного языка с их переводным эквивалентом на другом языке (или на нескольких других языках, в таком случае переводной словарь является многоязычным). Переводные словари условно разделяют на две большие группы:

• общелексические переводные словари. Переводят общую лексику с одного языка на другой или на несколько иностранных языков;

• научные, научно-технические и технические переводные словари. Включают в себя специальные термины по основным отраслям науки.

Несмотря на то, что существуют различные типы словарей, в их структуре можно выделить составные части, присутствующие под разными названиями практически во всех словарях. К таким элементам относятся:

• введение или предисловие (*Introduction*);

• раздел «Как пользоваться словарем» (*User's Mannual*);

• ключ к системе транскрипции, применяемой в словаре (*Keys to the Transcription*);

• список сокращений, используемых в словаре (*Contractions*);

• корпус словаря (*Corpus или The Body of the Dictionary*), то есть основной список слов, представленных их словарными статьями;

• дополнительный материал, то есть различные приложения (Ссорина, 2011).

Основу корпусного словаря составляют словарные статьи. Лингвистом Л.П.Ступиным в англо-русских переводных словарях выделены следующие части словарной статьи:

• *entryword / catchword / headword*— заглавное слово,

• *sense / meaningoftheword* — значение слова,

• *definition* — толкование, определение,

• *verbalillustration / quotation* — цитата, иллюстрация,

• *reference* — отсылка,

• *label* — метка,

• *status label* — метка о временной или территориальной ограниченности употребления слова,

• *regional label* — метка о территориальной употребительности слова,

• *functional label* — метка о принадлежности слова к части речи,

• *subject label* — метка о принадлежности слова к определенной области знаний (Ступин, 1985).

При разработке словаря для корпуса тувинского языка нами взята словарная статья со следующей структурой:

1) заглавное слово,

2) перевод (на русский язык, на английский язык и на другие языки),

3) транскрипция,

4) звучание заглавного слова,

5) метка о морфологических признаках (часть речи, число, падеж, склонение, спряжение),

6) значение слова,

7) этимологическая справка,

8) метка о принадлежности к аббревиатурам,

9) метка о наличии синонима, омонима и антонима,

10) дополнительная информация о слове.

База данных корпусного словаря включает следующие таблицы: MAIN — основная таблица с заголовочным словом; RUS, ENG, GER — таблицы с переводом заголовочного слова на различные языки; MORPHOLOGY — таблица с морфологическими данными. Структура этих таблиц представлена на рис. 1–4.

main

Имя поля	Тип данных	Описание (необя
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

Рисунок 1. Структура таблицы MAIN

Fig.1. MAIN table structure

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

Рисунок 2. Структура таблицы RUS

Fig.2 RUS table structure

eng

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

Рисунок 3. Структура таблицы ENG

Fig.3 ENG table structure

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

Рисунок 4. Структура таблицы MORPHOLOGY

Fig.4. MORPHOLOGY table structure

База данных реализована в Microsoft Office Access.

Для работы с корпусным словарем реализованы следующие функции: добавление новой статьи (рис. 5), редактирование статьи (рис. 6), удаление статьи (рис. 7), поиск словарной статьи с транскрипцией (рис. 8), формирование и визуализация морфологических признаков заглавного слова (рис. 9).

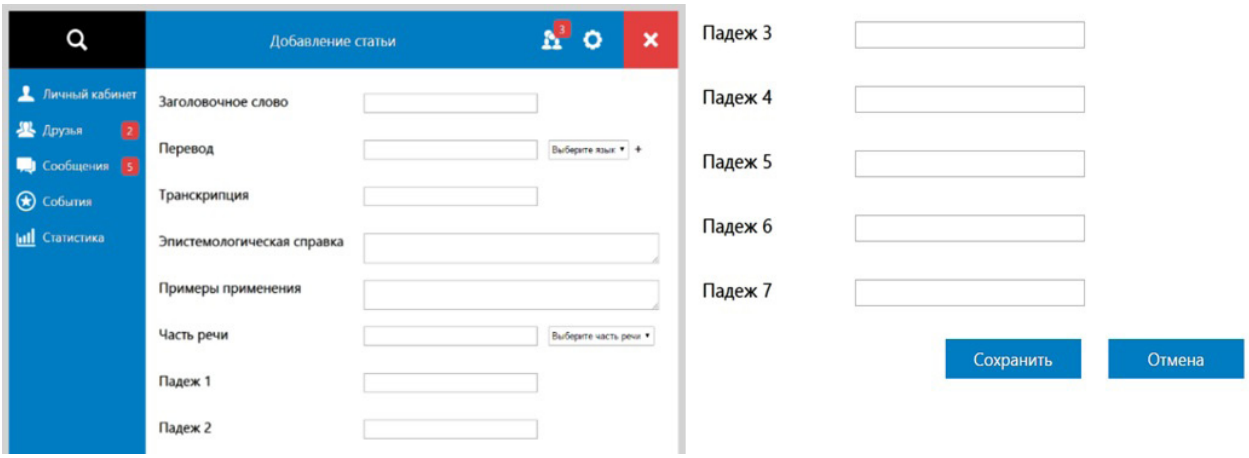

Рисунок 5. Интерфейс для добавления новой словарной статьи

Fig.5 Adding a new dictionary entry interface

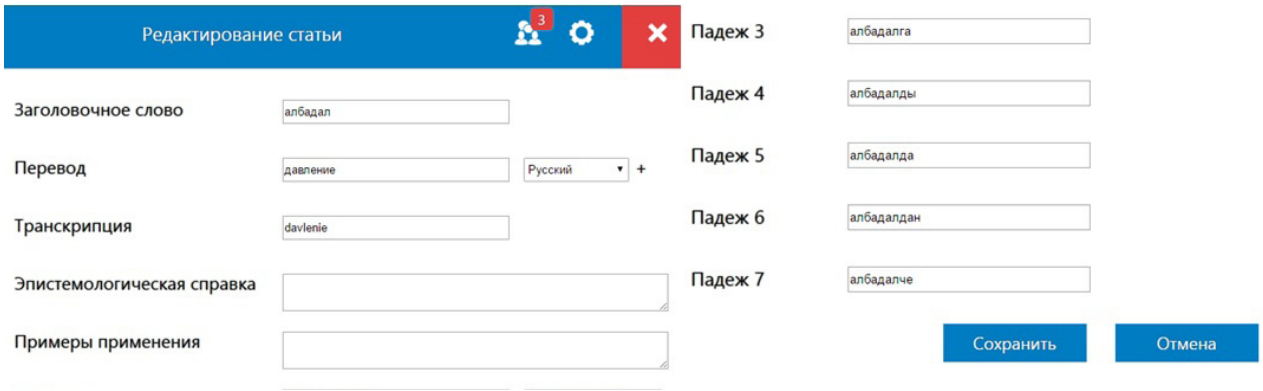

Рисунок 6. Интерфейс для редактирования статьи

Fig.6 Editing an entry interface

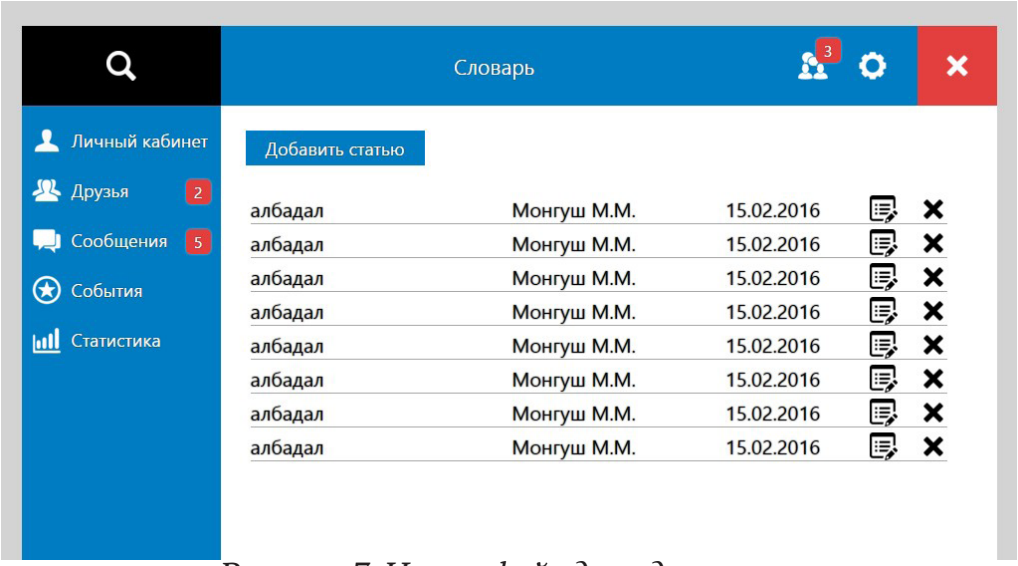

Рисунок 7. Интерфейс для удаления статьи

Fig.7. Entry removal interface

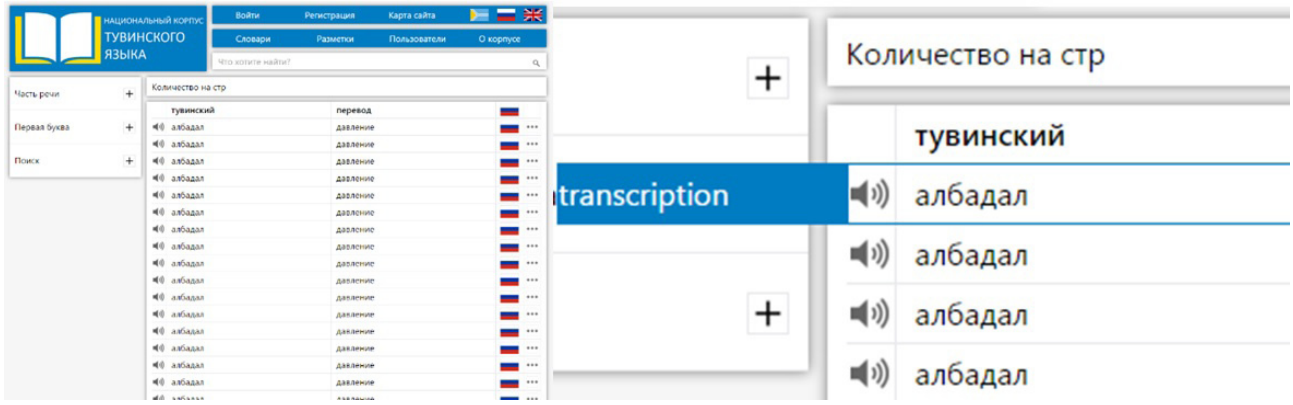

Рисунок 8. Поиск словарной статьи с транскрипцией

Fig.8. Entry search interface with transcription

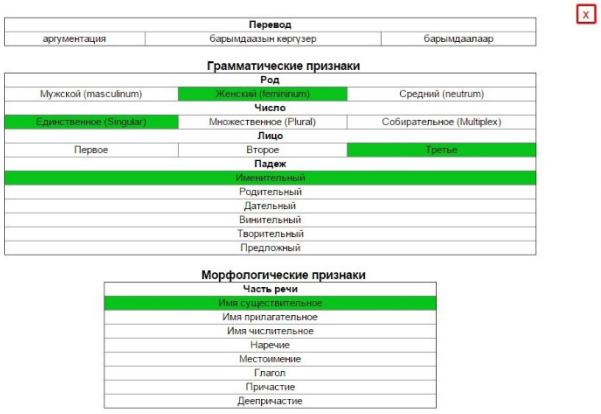

Рисунок 9. Морфологические признаки заглавного слова

Fig.9 Morphological features of a headword

Представленная выше структура словарной статьи и ее реализация в Microsoft Office Access позволяет рассматривать корпусный словарь как мультиструктурную организацию со сложным иерархическим строением, важнейшим корневым компонентом которого является словарная статья. Словарная статья является единицей корпусного словаря, которую можно формировать в различных аспектах и использовать для различных назначений: изучение тувинского языка с точки зрения написания, произношения, толкование, перевода на другие языки; формирование морфологической, синтаксической и семантической разметки текстов, входящих в корпус; играть роль тезауруса при анализе текстов; для организации поиска в корпусе.

Разработанный корпусный словарь может быть использован для изучения тувинского языка с точки зрения написания, произношения и толкования, а также для организации поиска слов и словосочетаний в текстах, хранящихся в корпусе.

В дальнейшем предполагается выполнение работ по формированию корпусного двуязычного словаря (тувинско-русского и русско-тувинского) с заполнением основных полей словарных статей преподавателями, аспирантами и студентами Тувинского государственного и Сибирского федерального университетов.

СПИСОК ЛИТЕРАТУРЫ

Бавуу-Сюрюн, М. В. (2010) Тувинский язык на современном этапе [Электронный ресурс] // Новые исследования Тувы. № 3. URL: http://www.tuva.asia/journal/issue_7/2158-bavyy-suyruyn-mv.html (дата обращения: 12.09.2016).

Салчак, А. Я., Байыр-оол, А. В. (2013) Электронный корпус тувинского языка: состояние, проблемы // Мир науки, культуры, образование. № 6. С. 408-409.

Ссорина, М. С. (2011) Словарь как мультиструктурная организация // Ярославский педагогический вестник. № 1. Т. 1. Гуманитарные науки. С. 142–146.

Ступин, Л. П. (1985) Лексикография английского языка : учебное пособие. М.: Высшая школа. 185 с.

Дата поступления: 20.10.2016 г.

REFERENCES

Bavuu-Siuriun, M. V. (2010) Tuvinskii iazyk na sovremennom etape. *Novye issledovaniia Tuvy*, no. 3 [online] Available at: http://www.tuva.asia/journal/issue_7/2158-bavyy-suyruyn-mv.html (access data: 12.09.2016). (In Russ.).

Salchak, A. Ia. and Baiyr-ool, A. V. (2013) Elektronnyi korpus tuvinskogo iazyka: sostoianie, problem. *Mir nauki, kul'tury, obrazovanie,* no. 6, pp. 408-409. (In Russ.).

Ssorina, M. S. (2011) Slovar' kak mul'tistrukturnaia organizatsiia. *Iaroslavskii pedagogicheskii vestnik*, no. 1, vol. 1. Gumanitarnye nauki, pp. 142–146. (In Russ.).

Stupin, L. P. (1985) *Leksikografiia angliiskogo iazyka* : uchebnoe posobie. Moscow, Vysshaia shkola. 185 p. (In Russ.).

Submission data: 20.10.2016.

Библиографическое описание статьи:

Дагбажык А. С. Структура словарной статьи Национального корпуса тувинского языка [Электронный ресурс] // Новые исследования Тувы. 2016, № 4. URL: http://nit.tuva.asia/nit/article/view/612 (дата обращения: дд.мм.гг.).

Citation:

Dagbazhyk A. S. The structure of an entry in the National corpus of Tuvan language. *Novye issledovaniia Tuvy*, 2016, no. 4 [on-line] Available at: http://nit.tuva.asia/nit/article/view/612 (accessed: ...).