ЭВМ И ТУВИНСКИЙ ЯЗЫК: ОБЗОР ИССЛЕДОВАТЕЛЬСКИХ РАБОТ ТУВИНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

COMPUTERS AND TUVAN LANGUAGE: AN OVERVIEW OF RESEARCH AT TUVAN STATE UNIVERSITY

Сергей М. Далаа
Тувинский государственный университет

Sergey M. Dalaa
Tuvan State University

Задача обработки текстовой информации поставлена и решается как в филологии, так и в информатике с самого зарождения этих наук. Особенно актуальной она стала с развитием Интернета. В Республике Тыва в настоящее время этой исследовательской работой занимается научно-образовательный центр «Тюркология» при ТувГУ совместно с кафедрой информатики этого же университета. В статье приведен обзор совместных исследований.

Сложность обработки текстов на тувинском языке с помощью компьютера заключается в том, что алфавит тувинского языка создан на основе алфавита русского языка (то есть русского варианта кириллицы) с добавлением трех букв ң, ө, ү. Это не вписывалось в стандарты международных кодов. Только с появлением в 1990-х гг. кодировки UNICODE, где тувинские буквы заняли свое место, эта проблема была частично решена. Появились системы программирования, которые поддерживают кодировку UNICODE.

Приведены аннотации баз данных и программ

Since their very beginnings, both philology and information technologies have faced the challenge of processing textual information. With the arrival of the Internet, this task has become more topical than ever before. In the Republic of Tuva, it is being dealt with by the Research and Education Center for Turkic Studies at the Tuvan State University in collaboration with the university's Department of Information Technologies. This article is an overview of their joint projects.

Computer processing of texts in Tuvan used to be a difficult task since the Tuvan alphabet is Cyrillic-based, but makes use of three letters absent in Russian - ң, ө and ү which did not have special codes assigned before the arrival of UNICODE. When the spread of UNICODE began in 1990s, Tuvan texts finally could be coded in their entirety.

The article provides short summaries and abstracts of databases and software created by Tuva's researchers in collaborative projects and registered at the Federal Service for Intellectual

Далаа Сергей Монгушевич — кандидат физико-математических наук, доцент кафедры информатики Тувинского государственного университета. Адрес: 667007, Россия, г. Кызыл, ул. Калинина, д. 30. Тел.: +7 (923) 262-72-54. Эл. адрес: dalaa@rambler.ru

Dalaa Sergey Mongushevich, Candidate of Physics and Mathematics, Associate Professor, Department of Information Technologies, Tuvan State University. Postal address: 30 Kalinin St., 667007 Kyzyl, Republic of Tuva, Russian Federation. Tel.: +7 (499) 374-75-95. E-mail: dalaa@rambler.ru

для ЭВМ, созданных учеными в совместных проектах и зарегистрированных в Федеральной службе по интеллектуальной собственности (Роспатент) в 2013-2015 гг. Правообладателем патентов является Тувинский государственный университет. В частности, это программы для ЭВМ: «Частотный словарь по художественным произведениям на тувинском языке», «Поиск слов в тексте на тувинском языке», «Тыва дыл. Сөзүглел. Практиктиг стилистика 10-11 класстарга өөредилге ному», «Лексика ландшафта Тувы», программа управления сайтом «Писатели Тувы», «Морфемно-орфографический словарь тувинского языка»; базы данных «Словарь диалектных слов алтайского диалекта тувинского языка», «Морфемно-орфографический словарь тувинского языка», «Аналитические скрепы тувинского языка».

С появлением новых мобильных устройств и разнообразием платформ, на которых они разрабатываются (Android, iOS, Windows Phone и др.), актуальной является разработка мобильных приложений различного назначения.

Ключевые слова: тувинский язык; ЭВМ; программа для ЭВМ; база данных; обработка текстов; язык программирования; Тувинский государственный университет; мобильное приложение

Property (Rospatent) in 2013-2015. All patent rights belong to Tuvan State University. The list includes such pieces of software as "Chastotnyi slovar po khudozhestvennym proizvedeniiam na tuvinskom iazyke" (Frequency dictionary of literary texts in Tuvan language), "Poisk slov v tekste na tuvinskom iazyke" (Word search in Tuvan texts), "Tyva dyl. Sөzүglel. Praktiktig stilistika 10-11 klasstarga өөredilge nomu" (Practical stylistics for 10th and 11th grades), "Leksika landshafta Tuvy" (The vocabulary of Tuvan landscape), CMS "Pisateli Tuvy" (The writers of Tuva), databases "Slovar' dialektnykh slov altaiskogo dialekta Tuvinskogo iazyka" (A vocabulary of the Altai dialect of Tuvan language), Morfemno-orfograficheskii slovar' Tuvinskogo iazyka" (Morphemic and orthographic dictionary of Tuvan language), and "Analiticheskie skrepy Tuvinskogo iazyka" (Analytical foundations of Tuvan language)

Given the rise of mobile devices and the variety of platforms for mobile software (Android, iOS, Windows Phone, etc.), development of mobile apps of various types is a highly topical challenge at the moment.

Keywords: Tuvan language; computer; software; database; text procession; programming language; Tuvan State University; mobile app

Введение

Задача обработки текстовой информации поставлена и решается как в филологии, так и в информатике с самого зарождения этих наук. Особенно актуальной она стала с развитием глобальной сети Интернет. В Республике Тыва обработкой текстов на тувинском языке с помощью ЭВМ впервые занялись исследователи из Тувинского государственного университета: доцент кафедры тувинского языка и литературы М. В. Бавуу-Сюрюн и доцент кафедры информатики О. Б. Бузур-оол (первый ректор Тувинского государственного университета) в 1980-х гг.

В настоящее время этой исследовательской работой занимается научно-образовательный центр (НОЦ) «Тюркология» при Тувинском государственном университете, которым руководит профессор М. В. Бавуу-Сюрюн, а со стороны кафедры информатики в этой работе активное участие принимает автор данной статьи (см.: Далаа, 2013).

В данной статье представлен обзор перспективных проектов, которые выполняются или запланированы к выполнению НОЦ «Тюркология» и кафедрой информатики в сфере информатизации проектов по развитию тувинского языка.

Направления исследований по обработке тувинских текстов

Сложность обработки текстов на тувинском языке с помощью компьютера заключается в том, что алфавит тувинского языка создан на основе алфавита русского языка (то есть русского варианта кириллицы) с добавлением трех букв ң, ө, ү (Поппе, 1929). Как мы знаем, кириллица была включена в международный стандарт ASCII, представляющий кодировки национальных алфавитов. Однако, поскольку в тувинском алфавите есть и свои буквы, то они не имели соответствия в ASCII. На первых порах создавались компьютерные шрифты, в которые добавлялись эти буквы, изменяя уже существующие символы этого шрифта. При этом не было единого стандарта на такие шрифты и эти шрифты нужно было устанавливать на каждый компьютер, чтобы работать с тувинскими текстами.

И только с появлением в начале 1990-х гг. кодировки UNICODE, где тувинские буквы заняли свое место, эта проблема была частично решена, так как эти буквы находятся в таблице UNICODE отдельно от букв русского алфавита. А это накладывает некоторые проблемы на создание алгоритмов обработки текстов на тувинском языке с помощью компьютера, а именно:

- усложняется код алгоритма;
- увеличивается время обработки текста (что является критичным для больших по объему текстов).

Появились системы программирования, которые поддерживают кодировку UNICODE. Например, языки программирования C#, PHP и другие. Это позволило создавать программы обработки текстов на тувинском языке.

Анализируя все работы по данной проблеме в Республике Тыва, можно выделить следующие направления исследований:

1. Создание баз данных (БД),
2. Создание алгоритмов обработки тувинского языка (АО),
3. Создание электронных учебников (ЭУ),
4. Создание локальных систем управления баз данных (ЛСУБД),
5. Создание сетевых систем управления баз данных (ССУБД).

Наряду с преподавателями Тувинского государственного университета в этой исследовательской работе также принимают участие студенты филологического и физико-математического факультетов. Тематика курсовых и дипломных работ студентов связываются с тематикой исследований НОЦ «Тюркология».

Можно привести примеры работ за последние годы (см. таб. 1).

Таблица 1. Исследовательские проекты НОЦ «Тюркология».

Table 1. Research projects at the Research and Education Center of Turkic Studies

Год	Название	Направление	Исполнитель	Руководитель от НОЦ "Тюркология"
2012	Аналитические скрепы тувинского языка	БД, ЛСУБД, ССУБД	Чыпсынак Ч. Ч.	Бавуу-Сюрюн М. В., Соян А. М.
2013	Словарь диалектных слов алтайского диалекта тувинского языка	БД, ЛСУБД, ССУБД	Саая Э. Э., Ооржак О.С.	Бавуу-Сюрюн М. В., Цэцэгдарь Уламсурэн, Хийс Гансух, Бадарч Баярсайхан
2013	Тыва дыл. Сөзүглел. Практиктиг стилистика 10-11 класстарга өөредилге ному.	ЭУ	Монгуш Ч.М.	Бавуу-Сюрюн М. В., Ооржак Л. Х.
2014	Морфемно-орфографический словарь тувинского языка	БД, ЛСУБД, АО	Ондар Ш. О.	Бавуу-Сюрюн М. В.
2015	Глагольные основы тувинского литературного языка	БД, ЛСУБД	Шмит Н. Н.	Хертек А. Б., Ооржак Б. Ч., Нара-Мандып А.А.
2015	Лексика ландшафта Тувы	БД, ЛСУБД	Дамба А. Г.	Бавуу-Сюрюн М. В.
2015	Программа управления сайтом «Писатели Тувы»	БД, ССУБД	Тумат С. К., Хуурак Ч. Б.	Бавуу-Сюрюн М. В.
2015	Автоматический поиск падежных аффиксов в текстах на тувинском языке средствами языка PHP	БД, ССУБД, АО	Тюлюш А. А.	Салчак А. Я., Байыр-оол А. В.

Все базы данных и программы для ЭВМ, созданные в вышеперечисленных студенческих работах, зарегистрированы в Федеральной службе по интеллектуальной собственности (Роспатент). Правообладателем патентов является Тувинский государственный университет.

Запатентованные исследовательские проекты

Среди совместных исследовательских проектов, в том числе получивших патенты Роспатента, можно назвать следующие. Дадим их названия и аннотации.

Проект «Программа для ЭВМ: Частотный словарь по художественным произ-

ведениям на тувинском языке» (авторы: С. М. Далаа, Т. А. Арапчор, М. В. Бавуу-Сюрюн, патент 10 сентября 2012 г. под №2012618172). Данная программа для ЭВМ предназначена для создания частотного словаря по художественным произведениям на тувинском языке на сайте «Электронный корпус тувинского языка» (tuvacorpus.ru) в глобальной сети Интернет. Она разбивает произведения на слова в алфавитном порядке тувинского языка и подсчитывает их количество повторений в данном произведении. Эта программа позволяет редактировать базу файлов художественных произведений на сервере (удалять и добавлять на сервер файлы). Доступ к базе возможен только по паролю. Программа создана с помощью свободно распространяемого языка программирования PHP версии 5.2.5.

База данных «Словарь диалектных слов алтайского диалекта тувинского языка» (авторы: М. В. Бавуу-Сюрюн, С. М. Далаа, Цэцэгдарь Уламсурэн, Бадарч Баярсайхан, Хийс Гансух, Э. С. Саая, О. С. Ооржак, патент от 12 сентября 2013г. №2013621145). Словарь диалектных слов алтайского диалекта тувинского языка составлен на основе полевых материалов, собранных составителями, с 1990-х гг. на территории Монголии и Китая в местах компактного проживания этнических тувинцев. Он отражает особенности алтайского диалекта тувинского языка, в то же время в нем отдельно помечены слова, характерные для отдельных говоров данного диалекта. Словарь будет представлять интерес для исследователей не только тувинского языка, но и тюркских языков Южной Сибири, а также монгольских языков и ученых, занимающихся типологическими исследованиям в области алтайской языковой общности.

Проект «Программа для ЭВМ: Тыва дыл. Сөзүглел. Практиктиг стилистика 10-11 класстарга өөредилге ному» (авторы М. В. Бавуу-Сюрюн, С. М. Далаа, Ч. М. Монгуш, Л. Х. Ооржак, патент от 28 августа 2013 г. №2013617990). Данная программа для ЭВМ представляет собой электронный учебник по тувинскому языку для 10-11 классов национальных школ Республики Тыва, который создан по одноименному печатному учебнику автора М. В. Бавуу-Сюрюн. Кроме того, в программу добавлены практическая часть (работа с рабочими тетрадями, созданными М. В. Бавуу-Сюрюн и Л. Х. Ооржак) и тестовая система, которая позволяет протестировать учащегося по теоретической части учебника. Эта программа позволяет также добавлять преподавателю в файл теста свои вопросы. Доступ к файлу теста защищен паролем. Программа создана с помощью объектно-ориентированного языка программирования Delphi 2009, который позволяет работать в кодировке UTF-8, что решает проблему вывода букв тувинского алфавита в интерфейсе программы.

Проект «Программа для ЭВМ: Поиск слов в тексте на тувинском языке» (авторы: М. В. Бавуу-Сюрюн, С. М. Далаа, Т. А. Арапчор, патент от 4 сентября 2013 г. под №2013618211). Данная программа для ЭВМ представляет собой макрос на языке программирования Visual Basic for Application (VBA), встроенного в офисный пакет Microsoft Office 2003. Макрос создан в текстовом файле фор-

мата doc. Эта программа разбивает заданный тувинский текст на страницы и находит данные слова и их словоформы на тувинском языке в этих страницах. Результатом поиска является частота данного слова (а также его словоформ) на отдельных страницах заданного текста и все это записывается в отдельный текстовый файл формата doc. Кроме этого, в файле результата создаются ссылки на страницы, по которым можно перейти и посмотреть в них найденные слова и их словоформы, выделенные красным цветом тексте. Кроме этого, имеется текстовый файл формата doc, в котором можно хранить слова для поиска, и этот файл можно редактировать с помощью программы.

Проект «База данных: "Морфемно-орфографический словарь тувинского языка"» (авторы: М. В. Бавуу-Сюрюн, С. М. Далаа, Ш. О. Ондар, патент от 3 декабря 2014 г. под №2015620151). Предлагаемый электронный словарь является комплексным: в нем даны не только основы слов, но и все словоформы, сложные с точки зрения их правописания. Все слова выстроены в алфавитном порядке. В отличие от орфографического словаря показана морфемная структура слова, в необходимых случаях указаны морфонологические процессы. Для последующего автоматизированного поиска нужных форм были введены спецсимволы.

Проект «Программа для ЭВМ: Лексика ландшафта Тувы» (авторы: М. В. Бавуу-Сюрюн, С. М. Далаа, А. Г. Дамба, патент от 18 сентября 2015 г. под №2015660008). Данная программа для ЭВМ представляет собой систему управления базой данных "Лексика ландшафта Тувы" в формате Microsoft SQL Server. Эта программа имеет два режима работы: пользовательский и администраторский. Для пользователя разрешен только просмотр и поиск данных в базе. Администратор базы данных может добавлять, изменять и удалять записи в базе. Администраторский доступ к базе защищен паролем. Программа создана с помощью объектно-ориентированного языка программирования C# из Visual Studio 2010, который позволяет работать в кодировке UTF-8, что решает проблему вывода букв тувинского алфавита в интерфейсе программы.

Проект «Программа для ЭВМ: Программа управления сайтом "Писатели Тувы"» (авторы: М. В. Бавуу-Сюрюн, С. М. Далаа, С. К. Тумат, Ч. Б. Хуурак, патент от 18 сентября 2015 г. под №2015660007). Данная программа для ЭВМ предназначена для управления сайтом "Писатели Тувы". Автором модели функционирования программы является М.В. Бавуу-Сюрюн. Эта программа имеет два режима работы: пользовательский и администраторский. Для пользователя разрешен только просмотр контента веб-сайта. Администратор сайта может добавлять, изменять и удалять содержимое сайта. Администраторский доступ к базе защищен паролем. Программа создана с помощью свободно распространяемых языков программирования PHP и JavaScript. Контент сайта хранится в базе данных формата MySQL.

Проект «Программа для ЭВМ: Морфемно-орфографический словарь тувин-

ского языка» (авторы: М. В. Бавуу-Сюрюн, С. М. Далаа, Ш. Ою Ондар, патент от 18 августа 2014 г. под №2015618798). Данная программа для ЭВМ представляет собой систему управления базой данных “Морфемно-орфографический словарь тувинского языка” в формате файла Microsoft ACCESS 2010. Автором базы является М. В. Бавуу-Сюрюн (патент от 3 декабря 2014 г. под №2015620151). Эта программа имеет два режима работы: пользовательский и администраторский. Для пользователя разрешен только просмотр и поиск данных в базе. Администратор базы данных может добавлять, изменять и удалять записи в базе. Администраторский доступ к базе защищен паролем. Кроме этого, программа производит поиск выбранных слов в словаре в текстовых файлах. Программа создана с помощью объектно-ориентированного языка программирования C# из Visual Studio 2010, который позволяет работать в кодировке UTF-8, что решает проблему вывода букв тувинского алфавита в интерфейсе программы.

Проект «База данных: “Аналитические скрепы тувинского языка”» (авторы: М. В. Бавуу-Сюрюн, С. М. Далаа, Ч. А. Бюрбю, Л. А. Шамина, А. М. Соян, патент от 24 октября 2012 г. под №2012621105). База данных содержит информацию о аналитических скрепах тувинского языка. Скрепы ¬¬ это сложные средства связи частей как сложного предложения, так и более крупных синтаксических построений. Они, как правило, многокомпонентны и строятся из делексикализованных слов с обязательным участием местоимений. Они в базе классифицированы по их функции на разных языковых уровнях, по выражаемым синтаксическим отношениям. А также разделены на внутрифразовые и межфразовые скрепы. Данная база является первым опытом представления и систематизации аналитических скреп по тюркским языкам Южной Сибири.

Заключение

Все эти названные проекты мы можем считать перспективными, в том числе для совместной работы. С появлением новых мобильных устройств и разнообразием платформ, на которых они разрабатываются (Android, iOS, Windows Phone и др.), актуальной является разработка мобильных приложений различного назначения. Мобильное приложение представляет собой программу, установленную на той или иной платформе, обладающую определенным функционалом, позволяющим выполнять различные действия.

В соответствии с этим одним из новых направлений исследований НОЦ «Тюркология» и кафедры информатики Тувинского государственного университета является разработка мобильных приложений с операционной системой Android. Этому способствуют следующие тенденции: к 2020 г. смартфоны практически полностью вытеснят с рынка обычные мобильные телефоны; смартфоны для многих людей становятся предметом первой необходимости; активная

политика продавцов по привлечению покупателей может сделать смартфоны более доступными (например, за счет субсидирования устройств операторами сотовой связи).

СПИСОК ЛИТЕРАТУРЫ

Далаа, С. М. (2013) Обработка текстов на тувинском языке средствами языка программирования PHP // Актуальные проблемы диалектологии языков народов России. Материалы XIII международной конференции. Уфа : Уфимский научный центр Академии наук России. С. 182-184.

Поппе, Н. Н. (1929) Заметки по фонетике танну-тувинского языка в связи с вопросом об алфавите // Культура и письменность Востока. Баку. Вып. IV. С. 49-61.

Дата поступления: 30.10.2016 г.

REFERENCES

Dalaa, S. M. (2013) Obrabotka tekstov na tuvinskom iazyke sredstvami iazyka programmirovaniia PHP. In: *Aktual'nye problemy dialektologii iazykov narodov Rossii. Materialy XIII mezhdunarodnoi konferentsii*. Ufa, Ufimskii nauchnyi tsentr Akademii nauk Rossii. Pp. 182-184. (In Russ.).

Poppe, N. N. (1929) Zametki po fonetike tannu-tuvinskogo iazyka v sviazi s voprosom ob alfavite. *Kul'tura i pis'mennost' Vostoka*. Baku. Vol. IV. Pp. 49-61. (In Russ.).

Submission date: 30.10.2016.

Библиографическое описание статьи:

Далаа С. М. ЭВМ и тувинский язык: обзор исследовательских работ Тувинского государственного университета [Электронный ресурс] // Новые исследования Тувы. 2016, № 4. URL: http://nit.tuva.asia/nit/article/view/611 (дата обращения: дд.мм.гг.).

Citation:

Dalaa S. M. Computers and Tuvan language: an overview of research at Tuvan State University. *Novye issledovaniia Tuvy*, 2016, no. 4 [on-line] Available at: http://nit.tuva.asia/nit/article/view/611 (accessed: ...).