Электронный информационный журнал
«НОВЫЕ ИССЛЕДОВАНИЯ ТУВЫ»
www.tuva.asia № 2 2012

ЭЛЕКТРОННЫЙ КОРПУС ТЕКСТОВ ТУВИНСКОГО ЯЗЫКА

А. Я. Салчак

Аннотация: Текст статьи был представлена на Международной научно-практической конференции, посвященной 100-летию со дня рождения «Народного академика» Владимира Михайловича Наделяева (г. Кызыл).

Статья подготовлена в рамках работы по проекту «Электронный корпус текстов тувинского языка»при поддержке РГНФ (грант 11-04-12073в).

Ключевые слова: Интернет, тувиноведение, обзор, тувинский язык, сайт, прикладная лингвистика

Одним из приоритетных направлений современной прикладной лингвистики является корпусная лингвистика, занимающаяся разработкой общих принципов построения и использования лингвистических корпусов с использованием компьютерных технологий. «Под лингвистическим, или языковым корпусом текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения лингвистических задач. В понятие «корпус текстов» входит также система управления текстовыми и лингвистическими данными, которую в последнее время чаще всего называют корпусным менеджером (или корпус-менеджером) (англ. corpus manager)» (Захаров, 2005: 3).

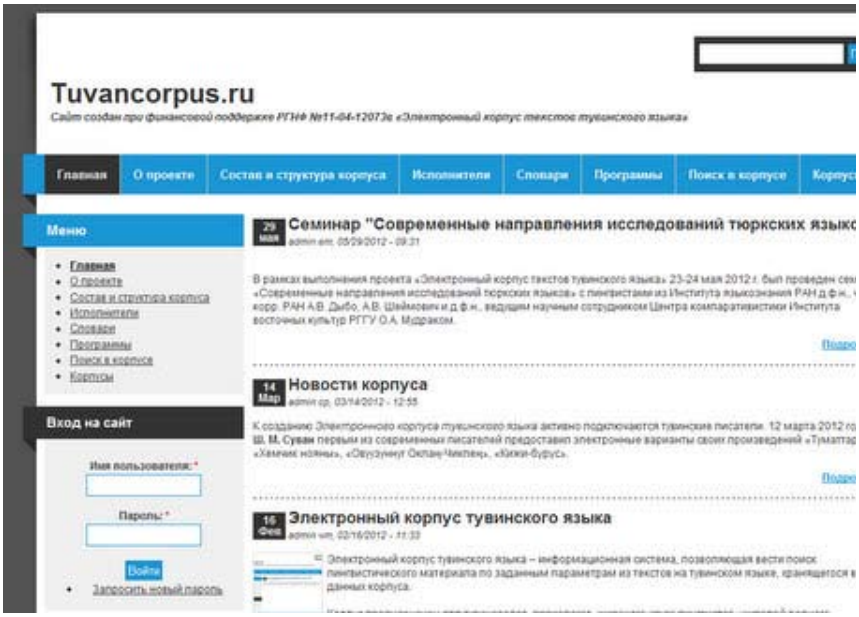
Первый большой корпус текстов на машинном носителе был создан в 1963 г. в Брауновском университете (США). Его создателями являются У.Френсис (W. Fransis) и Г. Кучера (H. Kucera). На сегодняшний день созданы корпусы для многих языков мира.

В апреле 2004 г. был открыт Национальный корпус русского языка (http://www.rus.corpora.ru). Тексты, представленные в национальном корпусе, имеют подробную лингвистическую и метатекстовую информацию.

Салчак Аэлита Яковлевна - кандидат филологических наук, старший преподаватель кафедры тувинской филологии и общего языкознания Тувинского государственного университета.

В национальном корпусе представлены тексты разных стилей, жанров и вариантов русского языка XIX и XX вв. Морфологический анализ в Национальном корпусе русского языка проводится с применением программ Mystem и Dialing.

Опыт создания электронных корпусов имеется в некоторых тюркских языках. Проект «Шорика», поддержанный Немецким научно-исследовательским обществом (ННИО) и Российским фондом фундаментальных исследований (РФФИ), был выполнен в течение 1999–2001 гг. В результате работы международной группы ученых был создан электронный корпус текстов по шорскому языку с применением четвертой версии программы Shoebox, разработанной с целью документации и помощи в развитии литературной формы бесписьменных и младописьменных языков народов мира. В корпусе шорского языка собраны литературные произведения (проза, стихи) и фольклорные тексты (эпос, сказки). Некоторые тексты сопровождаются транскрипцией на латинице, переводами на русский, английский языки и морфемной разбивкой. Имеется шорско-русский и русско-шорский онлайн-словарь (http://shoriya.ngpi.rdtc.ru/).



Разработана информационная система «Машинный фонд башкирского языка», представляющая собой автоматизированную информационную систему, состоящую из генеральной картотеки, лексикографической, эксперимен¬тально-фонетической, грамматической, диалектологической базы и каталога рукописных и старопечатных книг (http://mfbl.ru/). Лаборатория лингвистики и информационных технологий Института истории, языка и литературы Уфимского научного центра РАН приступает к разработке корпуса прозаических текстов башкирского языка, опубликованных с 40-х годов XX века по сегодняшний день. Сотрудники лаборатории приступили к работе над формированием перечня источников, оцифровкой

текстового материала, созданием систем экстралингвистических и внутрилингвистических разметок (Бускунбаева, Сиразитдинов, 2011: 50).

Продолжается работа над созданием Корпуса казахского языка (http://til.gov.kz). В рамках Программы фундаментальных исследований РАН «Корпусная лингвистика. Создание и развитие корпусных ресурсов по языкам народов России» ведется работа по созданию Корпуса хакасского языка (Шеймович, 2011: 48).

При поддержке Российского гуманитарного научного фонда в Научно-образовательном центре «Тюркология» Тувинского государственного университета с 2011 г. началась работа по созданию Корпуса тувинского языка (грант № 11-04-12073в «Электронный корпус текстов тувинского языка»).

Цель проекта — создание электронного корпуса тувинского языка, систематизированного собрания лингвистических банков данных, предназначенных для последующей комплексной автоматизации научных исследований и прикладных разработок в области тувинского языкознания, реализуемых на персональном компьютере. Основные задачи, которые предполагается решить, заключаются в создании:

1) базы данных тувинских текстов современного и советского периодов в локальной версии (формат базы данных Microsoft Access) и системы управления базой данных;

2) компьютерных программ для автоматизации сбора данных для лингвистических исследований в области тувинского языка (статистический метод);

3) базы данных частотных лексем и первичных именных и глагольных основ и системы управления базой данных (электронный словарь частотных лексем и первичных основ);

4) сайта электронного корпуса тувинских текстов.

На сегодняшний день переведены в электронный вид и отредактированы прозаические произведения тувинских писателей советского периода, основателей тувинской литературы С. А. Сарыг-оола, С. К. Токи, писателей второго поколения М. Б. Кенин-Лопсана, К.-Э. К. Кудажы, С. С. Сюрюн-оола, Е. Д. Тановой, а также писателей современного периода Э. Л. Донгака, Н. Ш. Куулара, З. С. Байсаловой, Ш. М. Сувана и некоторых других авторов. В Корпусе тувинского языка представлены некоторые поэтические тексты, фольклорные тексты, пьесы В. Ш. Кок-оола, С. К. Тока, В. С. Серен-оола, тексты официально-деловых документов на тувинском и русском языках

(Конституция Республики Тыва, и некоторые законодательные документы о выборах депутатов, должностных лиц и.т.д.), а также образцы фольклора и бытовой речи тувинцев Монголии.

В рамках проекта исполнителем проекта С. М. Далаа создана программа «Поиск морфем в заданном тексте» на языке программирования JavaScript, предназначенная для поиска морфем в текстах на тувинском языке.

Программа работает в браузере Internet Explorer с текстами, набранными в формате txt и кодировке UTF-8. Текст заранее набирается в файле (можно в программе Microsoft Word). Результатом программы является файл в в формате txt и кодировке UTF-8, который содержит все слова из заданного текста с выбранными морфемами. Программа успешно ищет словоформы с заданной морфемой, но возникли проблемы с омоформами. На данном этапе необходима ручная доводка соответствующей выборки. В дальнейшем программа будет расширена по своим функциональным возможностям.

Привлеченным в рамках проекта специалистом М. В. Бавуу-Сюрюн и исполнителем проекта С. М. Далаа подготовлен пробный морфемный словарь тувинского языка. 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. Омоформы рассматриваются каждый в отдельности. Для последующего автоматизированного поиска нужных форм были введены соответствующие знаки.

Совместно с привлеченным специалистом А. Ю. Серен был подготовлен пробный вариант электронного словаря с рабочим названием «ТывЛин». Текстовая часть электронного словаря была сделана исполнителем проекта А. В. Байыр-оол «ТывЛин» — разработанный по принципу электронных словарей ABBYY Lingvo электронный словарь, представляющий собой компьютерную программу, осуществляющий главным образом поиск тувинских слов с примерами и их переводов на русский язык. Эта программа предоставляет интуитивно понятный пользовательский интерфейс для перевода слов с тувинского языка на русский. Необходимое слово вводится с клавиатуры или находится из общего списка, и по нажатию клавиши «Enter» выводится его перевод.

На данный момент идет работа по составлению морфологических баз данных именных и глагольных форм тувинского языка. В базе данных представлены формы словоизменения и формообразования имен и глаголов. Разработана программа по составлению частотного словаря по выбранному произведению.

Все сделанное в рамках проекта на сегодняшний день доступно в сети Интернет по адресу http://www.tuvancorpus.ru

Список литературы:

Бускунбаева, Л. А., Сиразитдинов, З. А. (2011) К системе разметок в национальном корпусе башкирского языка // Актуальные проблемы диалектологии языков народов России. Материалы XI межрегиональной конференции. Уфа. С. 50–55.

Захаров, В. П. (2005) Корпусная лингвистика. СПб.

Шеймович, А. В. (2011) Морфологическая разметка корпуса хакасского языка // Российская тюркология. № 2(5). С. 48–61.

ELECTRONIC CORPS OF TEXTS OF TUVAN LANGUAGE

A. Ya. Salchak

Abstract: Article is prepared as a part of the project “Electronic corps of texts of Tuvan language” supported by the Russian Scientific Fund for the Humanities (grant #11-04-12073в).

Keywords: Internet, Tuvinology, review, Tuvan language, web-site, applied linguistics.