МЕТАТЕКСТОВАЯ РАЗМЕТКА В НАЦИОНАЛЬНОМ КОРПУСЕ ТУВИНСКОГО ЯЗЫКА: СТРУКТУРА И ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ

A METATEXTUAL MARKUP IN THE NATIONAL CORPUS OF TUVAN LANGUAGE: THE STRUCTURE AND FUNCTIONALITY

Чодураа М. Монгуш
Тувинский государственный университет,
Сибирский федеральный университет

Choduraa M. Mongush
Tuvan State University,
Siberian Federal University

Работа над формированием Национального корпуса тувинского языка (http://www.tuvacorpus.ru/) ведется преподавателями, аспирантами и студентами Тувинского государственного и Сибирского федерального университетов. В статье представлена система метаразметки корпуса, которая является важнейшей частью поискового аппарата всякого корпуса. Под метаразметкой понимается приписывание тексту параметров, характеризующих текст в целом. Метаразметка обеспечивает в корпусе возможность поиска и отбора текстов пользователем для составления подкорпусов с заданными свойствами. Отсюда следует, что чем больше набор признаков, по которым характеризуется каждый текст, тем шире возможности поиска текстов для решения различных лингвистических и филологических задач.

Creating natural language corpora helps solve a number of philological and purely linguistic problems for many languages of the peoples of Russian Federation. National corpus of Tuvan language (http://www.tuvacorpus.ru/) is one of such products jointly developed by faculty and students at two universities in Krasnoyarsk and Kyzyl.

The article presents a meta-markup system which forms the most important part of the search functionality in any corpus. Meta-markup refers to assigning parameters characterizing the text as a whole. Within a corpus, meta-markup provides the opportunity to search and select texts to include them into subcorpora by the presence of a certain feature(s). Consequently, the larger the set of such features is for each text, the wider become the search functionality for various philological and

Монгуш Чодураа Михайловна — преподаватель кафедры информатики Тувинского государственного университета; аспирант Института математики и фундаментальной информатики Сибирского федерального университета (г. Красноярск). Адрес: 667000, Россия, г. Кызыл, ул.Колхозная, д. 125. Тел.: +7 (929) 335-45-46. Эл. адрес: mongushchod@yandex.ru. Научный руководитель — д-р физ.-мат. н., проф. В. В. Быкова.

Mongush Choduraa Mikhailovna, Instructor, Department of Information Technologies, Tuvan State University; Postgraduate student, Institute of Mathematics and Information Technologies, Siberian Federal University. Postal address: 125 Kolkhoznaya St., 667000 Kyzyl, Republic of Tuva, Russian Federation. Tel.: +7 (929) 335-45-46. E-mail: mongushchod@yandex.ru. Research advisor: Doctor of Physics and Mathematics, Professor V.V. Bykova.

Система метаразметки текстов Национального корпуса тувинского языка может содержать 18 параметров, в том числе: имя автора, пол автора, название текста, дата создания текста — год написания текста, сфера функционирования текста, тема текста или предметная область, хронотоп, или место и время описываемых событий, принадлежность текста к определенному речевому жанру, жанр художественной литературы, стиль текста, возраст аудитории, уровень образования аудитории, источник текста, название издания, название издательства, год издания, тип носителя, комментарии.

Ключевые слова: корпусы естественных языков; Национальный корпус тувинского языка; метаразметка; тувинский язык; тувинский героический эпос

The meta-markup system for the texts included into the National corpus of Tuvan language may include up to 18 parameters, such as the author's name and gender, the title and creation date (year) of the text, its functional sphere, topic, subject area, time and setting of events described in it, the text's classification by type of spoken language or literary genre and style, its source, name of the periodical it appeared in, publisher, publication date, medium, comments, as well as some features of its audience, such as age and education level.

Keywords: natural language corpus; National corpus of Tuvan language; meta-markup; Tuvan language; Tuvan epic poetry

Введение

Национальный корпус тувинского языка создан в 2011 г. сотрудниками научно-образовательного центра «Тюркология» и преподавателями Тувинского государственного университета. К работе по переводу в электронный формат текстов и публицистики были подключены учителя тувинского языка и литературы ряда общеобразовательных школ Республики Тыва (Салчак, Байыр-оол, 2013). На сегодняшний день в Национальном корпусе тувинского языка, доступном в Интернете по адресу http://www.tuvacorpus.ru/, 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 «Тыв-Лин», словарь диалектных слов алтайского диалекта тувинского языка, морфемно-орфографический словарь тувинского языка, составленный М. В. Бавуу-Сюрюн и С. М. Далаа. В корпусе предусмотрен поиск слов и морфем в заданном тексте (Хертек, Ооржак, 2012).

Работы по расширению информационного содержания Национального корпуса и углублению уровня обработки текстов продолжаются, поскольку вопрос о сохранении тувинского языка как одной из основ целостности тувинского этноса и его национальной культуры не просто сохраняет остроту, проблемы продолжают углубляться (см.: Бавуу-Сюрюн, 2010: Электр. ресурс).

В данной статье представлена система метаразметки корпуса, которая является важнейшей частью поискового аппарата всякого корпуса.

Корпусы естественных языков и их назначение

Под корпусом понимается информационно-справочная система, основанная на собрании оцифрованных текстов. На основании корпусов решаются различные филологические и лингвистические задачи. Для многих языков народов Российской Федерации, в том числе для тюркских языков, создаются национальные корпусы.

Национальный корпус предназначен в первую очередь для лингвистов. Он обеспечивает возможность проведения научных исследований для решения различных лингвистических задач. Однако статистические данные о языке определенной эпохи или определенного автора интересны литературоведам, историкам и представителям многих других областей гуманитарного знания. Национальный корпус важен также для преподавания языка в качестве родного или иностранного (Сысоев, 2010). Анализ и обработка разных типов корпусов являются предметом большинства работ в области компьютерной лингвистики, распознавания речи и машинного перевода.

Всякий корпус включает в себя информационные и программные составляющие. Создание корпуса предполагает выполнение следующих работ:

1) определение перечня хранимых текстов,

2) оцифровка текстов,

3) выверка и корректировка текстов,

4) выбор типов разметки, способов их машинного представления и организации данных в целом,

5) разметка текстов (вручную или автоматически),

6) определение и реализация поискового аппарата — множества возможных запросов к данным, хранящимся в корпусе,

7) разработка программных средств обеспечения доступа к корпусу, т. е. интерфейса (Захаров, 2005).

Одним из основных этапов создания корпуса является формирование его информационной составляющей, т. е. выполнение работ 1–5, согласно перечню В.П.Захарова. Вопрос представительности информационной составляющей корпуса чрезвычайно важен. Перечень текстов, подлежащих хранению в корпусе, призван адекватно отражать лексико-грамматические феномены, типичные для исследуемого класса текстов соответствующего языка. Для представительности важен как объем хранимых данных, так и их организация. Эти параметры корпуса зависят от потенциальных запросов к хранимым данным. Множество возможных запросов характеризует функциональные возможности любого корпуса.

Оцифровка текстов — преобразование текстов в электронный вид, которое может быть выполнено самыми разными способами (ручной ввод, сканирование и распознавание текстов и т. п.). После оцифровки текстов необходимо провести их предобработку: выверку и корректировку текста. Разметка текста — приписывание текстам дополнительной информации. Хорошая разметка позволяет быстро и эффективно найти в корпусе слова, грамматические формы слов и конструкции, которые нужны исследователю. Традиционно в корпусах используется пять типов разметки: метатекстовая, морфологическая, синтаксическая, семантическая и др. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса (там же).

Морфологическая разметка текстов в корпусе представляет собой выделение каждой словоформы и приписывание к ней информации об ее принадлежности какой-либо части речи, определение ее грамматической формы в некотором контексте. Такая разметка необходима для того, чтобы осуществлять поиск нужных пользователю слов, форм слова, конструкций.

Синтаксическая разметка — это отражение синтаксической структуры предложения в тексте. Синтаксическая структура предложения представляет собой дерево зависимостей, в узлах которого стоят слова предложения, а ветви помечены именами синтаксических отношений. В синтаксической разметке существует два уровня: аналитический, на котором размечена поверхностная структура предложения (подлежащее, сказуемое, союзы, предлоги, знаки препинания и др.), и текстограмматический, отражающий глубинную структуру предложения.

Для русскоязычных текстов существует несколько лингвистических программ для разметки текстов: Russian Morphological Dictionary — программа для синтаксического и морфологического анализа русскоязычных текстов, Mistem ¬¬— компактный и очень быстрый морфологический парсер, реализованный на основе словаря Зализняка. Данные программы находятся в открытом доступе и являются бесплатными (см. использование таковой для Национального корпуса калмыцкого языка: Куканова, 2015: Электр. ресурс). Но надо сказать, что для автоматической разметки тувинских текстов эти программы не подходят.

Метаразметка тувинских текстов

Существенной частью корпуса является метаразметка. Под метаразметкой понимается набор признаков, характеризующий текст в целом. Совокупность этих параметров называют также паспортом текста. Метаразметка обеспечивает в корпусе возможность поиска и отбора текстов пользователем для составления подкорпусов с заданными свойствами (Савчук, 2005: Электр. ресурс). Отсюда следует, что чем больше набор признаков, по которым характеризуется каждый текст, тем шире возможности поиска текстов для решения различных лингвистических и

филологических задач, сводимых к задаче классификации и кластеризации.

Для описания текстов в базе данных Национального корпуса русского языка (НКРЯ) используется 25 признаков: 9 параметров характеризует текст, 3 параметра характеризует автора, 3 — возможную аудиторию, 4 параметра содержат библиографические данные о тексте, 5 параметров представляют собой служебную информацию. Подробное описание этих параметров представлены в работе С.О.Савчука (там же).

На основе этой системы метаразметки можно выделить следующие параметры, которые могут описывать тексты на тувинском языке:

• имя автора. Этот параметр включает несколько значений (конкретный автор, обобщенный автор, коллективный автор, неизвестный автор);

• пол автора — мужской, женский и неизвестен;

• название текста;

• дата создание текста — год написание текста;

• сфера функционирования текста — это сферы речевой деятельности. Выделяются следующие функциональные сферы: учебно-научная, производственно-техническая, официально-деловая, публицистики, рекламы, церковно-богословская, художественная, бытовая;

• тема текста или предметная область. В перечне тематических областей, предлагаемых одним из международных стандартов корпусной лингвистики EAGLES (там же; EAGLES… , Электр. ресурс), в одном ряду встречаются области, находящиеся в отношении соподчинения, например «естественные науки» и «физика», «математика», «биология» и пр. Поэтому Синклер предлагает выстроить их в виде раскрывающего списка (например, 2 и 2.1, 2.2, 2.3 …, 3 и 3.1, 3.2 и т. д.);

• хронотоп, или место и время описываемых событий;

• тип текста определяет принадлежность текста к определенному речевому жанру;

• жанр художественной литературы;

• стиль текста;

• возраст аудитории: детская (0–10 лет), подростковая (11–17 лет), молодежная (18–34 года), взрослая аудитория, или не оказывает существенного влияния на свойства текста;

• уровень образования аудитории. Выделяют 4 значения этого параметра: «высокий» — если текст рассчитан на читателя с высоким уровнем общего образования и с общим знанием о предмете, «профессиональный» — текст рассчитан на специалистов с различным уровнем общего образования, «низкий» — текст предназначен для нетребовательного читателя, «н-уровень» — в других случаях;

- источник текста;
- название издания;
- название издательства;
- год издания;
- тип носителя;
- комментарии.

Таковой в целом мы видим систему метаразметки тувинских текстов в Национальном корпусе тувинского языка с 18 параметрами. Данная система метаразметки текстов в Национальном корпусе тувинского языка позволит пользователю отбирать тексты по любому из признаков или их комбинациям и формировать подкорпус текстов для решения конкретных лингвистических задач.

По предложенной системе в Национальном корпусе тувинского языка были размечены тексты тувинского героического эпоса, и была создана база данных «Тувинские героические сказания», которая содержит паспортизацию тувинских героических сказаний и информацию о сказителях.

Заключение

На основе рассматриваемой базы данных можно разрабатывать различные компьютерные программы для выявления языковых особенностей тувинских героических эпосов, которые позволят увидеть национально-культурную специфику языковой картины мира кочевого народа. База данных по тувинским героическим сказаниям предназначена для комплексной автоматизации научных исследований и прикладных разработок в области тувинского языкознания, реализуемых на персональном компьютере.

СПИСОК ЛИТЕРАТУРЫ

Бавуу-Сюрюн, М. В. (2010) Тувинский язык на современном этапе: образовательный аспект [Электронный ресурс] // Новые исследования Тувы. № 3. URL: http://www.tuva.asia/journal/issue_7/2158-bavyy-suyruyn-mv.html (дата обращения: 12.06.2016).

Захаров, В. П. (2005) Корпусная лингвистика: учебно-методическое пособие . СПб. : БВХ-Петербург. 48 с.

Куканова, В. В. (2015) Национальный корпус калмыцкого языка: итоги работы и перспективы [Электронный ресурс] // Новые исследования Тувы. № 1. URL: http://www.tuva.asia/journal/issue_25/7760-kukanova.html (дата обращения: 01.04.2016).

Савчук, С. О. (2005) Метатекстовая разметка в национальном корпусе русского языка: базовые принципы и основные функции [Электронный ресурс] // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. URL: http://ruscorpora.ru/sbornik2005/05savchuk.pdf (дата обращения: 01.04.2016).

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

Сысоев, П. В. (2010) Лингвистический корпус в методике обучения иностранным языкам // Язык и культура. № 1. С. 99-111.

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

EAGLES. Text Corpora Working Group Reading Guide. EAG--TCWG--FR—2. Version of May, 1996 [Электронный документ] // Istituto di Linguistica Computazionale «A. Zampolli». URL: http://www.ilc.cnr.it/EAGLES96/corpintr/corpintr.html (дата обращения: 12.09.2016).

Дата поступления: 01.11.2016 г.

REFERENCES

Bavuu-Siuriun, M. V. (2010) Tuvinskii iazyk na sovremennom etape: obrazovatel'nyi aspekt. *Novye issledovaniia Tuvy*, no. 3 [online] Available at: http://www.tuva.asia/journal/issue_7/2158-bavyy-suyruyn-mv.html (access data: 12.06.2016). (In Russ.).

Zakharov, V. P. (2005) *Korpusnaia lingvistika*: uchebno-metodicheskoe posobie. St. Petersburg, BVKh-Peterburg. 48 p. (In Russ.).

Kukanova, V. V. (2015) Natsional'nyi korpus kalmytskogo iazyka: itogi raboty i perspektivy. *Novye issledovaniia Tuvy*, no. 1 [online] Available at: http://www.tuva.asia/journal/issue_25/7760-kukanova.html (access data: 01.04.2016). (In Russ.).

Savchuk, S. O. (2005) Metatekstovaia razmetka v natsional'nom korpuse russkogo iazyka: bazovye printsipy i osnovnye funktsii. *Natsional'nyi korpus russkogo iazyka: 2003–2005. Rezul'taty i perspektivy* [online] Available at: http://ruscorpora.ru/sbornik2005/05savchuk.pdf (access data: 01.04.2016). (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.).

Sysoev, P. V. (2010) Lingvisticheskii korpus v metodike obucheniia inostrannym iazykam. *Iazyk i kul'tura*, no. 1, pp. 99-111. (In Russ.).

Khertek, A. B. and Oorzhak A. B. (2012) O morfologicheskoi razmetke elektronnogo

korpusa tekstov tuvinskogo iazyka. *Gramota. Filologicheskie nauki. Voprosy teorii i praktiki*, no. 7, pp. 214–218. (In Russ.).

EAGLES. Text Corpora Working Group Reading Guide. EAG-TCWG-FR-2. Version of May, 1996. *Istituto di Linguistica Computazionale «A. Zampolli»* [online] Available at: http://www.ilc.cnr.it/EAGLES96/corpintr/corpintr.html (access data: 12.09.2016).

Submission data: 01.11.2016.

Библиографическое описание статьи:

Монгуш Ч. М. Метатекстовая разметка в Национальном корпусе тувинского языка: структура и функциональные возможности [Электронный ресурс] // Новые исследования Тувы. 2016, № 4. URL: http://nit.tuva.asia/nit/article/view/613 (дата обращения: дд.мм.гг.).

Citation:

Mongush Ch. M. A metatextual markup in the national corpus of Tuvan language: the structure and functionality. *Novye issledovaniia Tuvy*, 2016, no. 4 [on-line] Available at: http://nit.tuva.asia/nit/article/view/613 (accessed: ...).