УДК 519.7
DOI: 10.15827/2311-6749.17.3.2

БАЗА ДАННЫХ И СРЕДСТВА СОЗДАНИЯ КОНТЕКСТОВ ДЛЯ ПРЕДСТАВЛЕНИЯ И АНАЛИЗА ТУВИНСКОГО ГЕРОИЧЕСКОГО ЭПОСА

(Работа выполнена при поддержке Российского гуманитарного научного фонда, грант № 16-34-1-01033)

Ч.М. Монгуш, аспирант, mongushchod91@yandex.ru
(Сибирский федеральный университет, просп. Свободный, 79, г. Красноярск, 660041, Россия);
М.В. Ондар, аспирант, mengi89@yandex.ru
(Тувинский государственный университет, ул. Ленина, 36, г. Кызыл, 667000, Россия)

Национальный корпус тувинского языка содержит специальный раздел, посвященный тувинской литературе. В работе представлена БД фольклорного раздела корпуса тувинского языка, содержащая описание более 50 тувинских эпических сказаний. Предложена структура метаразметки текстов тувинского героического эпоса. Приведено описание программных средств формирования бинарных контекстов (матриц вида «объект–свойство») на основе БД корпуса. Представленные результаты могут быть использованы для изучения и анализа тувинского фольклора с помощью математических методов и современных информационных технологий.

Ключевые слова: *тувинский героический эпос, БД, метаразметка, бинарный контекст.*

В настоящее время с помощью современных информационных технологий и математического моделирования активно создаются корпусы естественных языков. 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. Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. С помощью корпусов решаются многие филологические и лингвистические задачи.

Работа над формированием Национального корпуса тувинского языка ведется преподавателями, аспирантами и студентами Тувинского государственного и Сибирского федерального университетов. Информационная составляющая корпуса активно пополняется учеными Тувинского государственного университета. В корпус включен специальный раздел, посвященный тувинской художественной литературе и фольклору. В него также входят частотный словарь по художественным произведениям на тувинском языке, тувинско-русский электронный словарь «ТывЛин», словарь диалектных слов алтайского диалекта тувинского языка, морфемно-орфографический словарь [1]. В корпусе предусмотрен поиск слов и морфем в заданном тексте. Работы по расширению информационного содержания Национального корпуса и углублению уровня обработки текстов продолжаются [2]. Особый интерес представляют исследования произведений тувинского героического эпоса как важной составляющей этнокультурного наследия Республики Тыва [3].

Всякий корпус как информационно-справочная система включает в себя информационную и программную составляющие. Создание корпуса предполагает выполнение следующих работ: определение перечня хранимых текстов, оцифровка текстов, выверка и корректировка текстов, выбор типов разметки, разметка текстов (вручную или автоматически), разработка программных средств обеспечения доступа к хранимым текстам.

Существенной частью поискового аппарата корпуса является метаразметка. Под метаразметкой понимается приписывание тексту множества параметров, значения которых характеризуют текст в целом. Совокупность значений этих параметров называется метаописанием, или паспортом произведения. Определение структуры метаописаний и формирование метаописаний – важнейшая задача, возникающая при создании корпуса. Информация, отражающая метаописания исследуемого множества текстов, в большинстве случаев хорошо структурирована и допускает представление контекстом – матрицей вида «объект–признак», каждая строка которой содержит метаописание конкретного текста. Такое представление информации о текстах позволяет применять при их исследовании математический аппарат машинного обучения и анализа формальных понятий и решать различные лингвистические и филологические задачи, сводимые к концептуальному моделированию и классификации по прецедентам [4, 5]. Определение состава признаков, входящих в метаописание, является достаточно сложной задачей, требующей привлечения лингвистов и филологов. Состав признаков, как правило, устанавливается, исходя из задач исследования, на которые ориентирован корпус.

В данной статье представлена БД фольклорного раздела корпуса тувинского языка, содержащая описания более 50 тувинских эпических сказаний. Предложена структура метаразметки текстов тувинского

героического эпоса, приведено описание программных средств формирования бинарных контекстов на основе БД корпуса.

БД тувинских героических сказаний

Богатый фонд рукописных и магнитофонных записей всех жанров тувинского фольклора находится в научном архиве *Тувинского института гуманитарных и прикладных социально-экономических исследований* (ТИГПИ). Основу архива составляют полевые фольклорно-этнографические материалы, собранные в Республике Тыва и у этнических тувинцев Китая и Монголии. Исключительная роль по введению в научный оборот текстов тувинских героических сказаний принадлежит коллективу ТИГПИ. В архиве института хранятся около 300 записей эпических произведений [3]. Однако свет увидели немногие, поскольку напечатаны они в очень старых ветхих книгах. В настоящее время имеются 14 сборников и отдельных изданий, содержащих тувинские героические сказания.

В БД Национального корпуса тувинского языка включены тексты тувинского героического эпоса, представленные в [3]. Сведения о сборниках, в которых опубликованы эти тексты, и названия этих текстов приведены в таблице 1. В настоящее время тексты тувинских сказаний продолжают расшифровываться и издаваться сотрудниками ТИГПИ. Поэтому сформированная БД будет периодически пополняться новыми фольклорными произведениями, в том числе и за счет фольклора тувинцев зарубежья.

В БД хранятся не только оцифрованные тексты произведений, но и их метаописания. Набор признаков, составляющих метаописание текста произведения, считается релевантным, если эти признаки отражают текст с существенной для исследователя точки зрения. Релевантный набор признаков выбирается экспертами в зависимости от филологических и лингвистических задач, решаемых в рамках корпусов.

Применительно к тувинскому героическому эпосу были определены релевантные наборы признаков, определяющие метаописания текстов героических сказаний тувинского народа. Данные наборы признаков были согласованы с сотрудниками ТИГПИ и профессором Тувинского государственного университета, директором научно-образовательного центра «Тюркология» М.В. Бавуу-Сюрюн. Установленные наборы признаков позволяют формировать различные контексты произведений тувинского героического эпоса и применять при их изучении и исследовании современные математические методы [7, 8].

В состав метаописания текстов тувинского героического эпоса входят род, вид, сюжет, мотив, стандартные словоупотребления или клише, форма, герой, зачин произведения. Название и год издания сборника, в котором опубликован текст, также входят в метаописание. Многократное вхождение текста в различные сборники обязательно фиксируется в метаописании. Так, было установлено, что из указанных в таблице 1 текстов в разных изданиях встречаются следующие произведения.

1) «Алдын-Кургулдай», сказитель – Догаа Соян Куутсулмааевич из Эрзинского района (опубл. в двух сборниках, изданных в 1957 и 1993 гг.).

2) «Анан-Даваа», сказитель – Саая Одербей Мызаа-Каракович из Чеен- Хемчикского района (опубл. в сборниках 2012 и 2014 гг.).

3) «Арзылан-Кара аттыгЧечен-Кара меге», сказитель – Дондук Салчак Дамдынович из Бай-Тайгинского района (опубл. в сборниках 1995 и 2012 гг.).

4) «Арзылан-Кара аттыгХунан-Кара», сказитель – Ооржак Чанчы-Хее Чапаажыкович (опубл. в сборниках, вышедших в 1997 г. и 2014 гг.)

5) «0лээдей-Мерген», сказитель – Хертек Шой Чамзыевич из Монгун-Тайгинского района (опубл. в сборниках 2012 и 2014 гг.).

6) «Эрелзей-Мерген, Харагалзай-Мерген алышкылар», сказитель – Дамдын Оюн Хорлуу из Тандинского района (опубл. в сборниках, вышедших в 1955 и 1993 гг.).

Для БД корпуса тувинского языка существенен также вопрос о наличии переводов тувинских фольклорных произведений на русский язык. С этой целью в метаописание текстов введен параметр, отражающий наличие перевода. С помощью данного параметра возможно формирование параллельного корпуса переводов тувинского героического эпоса.

Для научных исследований, проводимых в лингвофольклористике, чрезвычайно важны варианты одного и того же текста, исполненные разными сказителями. Это требуется для распознавания индивидуального стиля сказителя. Необходимые параметры, предназначенные для отражения данных сведений, также введены в метаописания текстов, хранимых в БД. Например, были выявлены четыре варианта сказания «Бокту-Кириш, Бора-Шээлей», записанного в разное время разными сказителями и в разных районах Тувы. Список этих вариантов с указанием значений соответствующих параметров метаописаний приведен в таблице 2.

Таким образом, метаописание каждого текста героического эпоса в БД Национального корпуса тувинского языка включает в себя информацию о записи текста (место записи, кем и в каком году записан, в каком сборнике издан), о сказителе (Ф.И.О., год и место его рождения, статус по грамотности, место рождения и проживания), о сборниках, где опубликован текст (название, составитель, в каком году и где он издан), об имеющихся переводах текста на другие языки. БД тувинских героических сказаний создана

на основе средств Microsoft Office Access и запатентована [9]. Она представляет собой этнокультурную ценность и служит основой для создания бинарных контекстов с последующим применением к ним математического аппарата машинного обучения и анализа формальных понятий.

Таблица 1

Перечень использованных сборников тувинских героических сказаний

Название сборника	Составители	Год издания	№ текста	Название текста
Тыва тоолдар	А.К. Калзан	1955	1	Эрелзей-Мерген, Харагалзай-Мерген алышкылар
Тыва тоолдар	О. Сувакпит	1957	2 3 4 5 6 7	Демир-Шилги аъттыг Тевене моте Алдын-Кургулдай Карыш-Кулаш хаайлыг Калчан-Шилги аъттыг Куре-Хевек Хартыга-Бора аъттыг Чащыс-Карыш Ээр-Сарыг аъттыг Экер-оол Уран, Маадыр ийи угбашкы
Арзылац-Мерген	Ч.Ч. Куулар, О. К.-Ч. Дарыма	1974	8 9	Кучун-Хурец аъттыг Хуру-Маадыр Чеди харлыг Хан-Тегулдур
Тыва маадырлыг тоолдар, I том	С.М. Орус-оол	1990	10 11 12 13 14	Меге Шагаан-Тоолай Танаа-Херел Тоц-Аралчын-Хаан Кацгывай-Мерген Баян-Тоолай
Алдай-Буучу, II том	С.М. Орус-оол	1993	15 16 17 18 19	Алдай-Буучу Алдын-Чаагай Бокту-Кириш, Бора-Шээлей Алдын-Кургулдай Эрелзей-Мерген, Харагалазай-Мерген алышкылар
Далай-Байбыц-Хаан, III том	С.М. Орус-оол	1994	20 21 22 23 24 25	Найгы-Майгы маадыр Коцгар-Баадай Хаан-Тегулдур Далай-Байбыц-Хаан Хеекуй-Кара Бораадай-Мерген
Боктуг-Кириш, Бора-Шээлей, IV том	С.М. Орус-оол	1995	26	Боктуг-Кириш, Бора-Шээлей
Ары-Хаан, V том	С.М. Орус-оол	1996	27 28 29 30	Хан-Шилги аъттыг Хан-Хулук Ары-Хаан Арзылац-Кара аъттыг Чечен-Кара меге Шеегун-Бора аъттыг Шеегун-Кеегун
Тувинские героические сказания. Памятники фольклора народов Сибири и Дальнего Востока	С.М. Орус-оол	1997	31 32	Хунан-Кара Боктуг-Кириш, Бора-Шээлей
Тыва улустуц тоолдары (Тываныц тоолчуларыныц 2 дугаар следунуц материалдары)	С.М. Орус-оол	2012	33 34 35 36 37 38 39 40 41 42	Арзылац-Кара аъттыг Чечен-кара меге Эртине-Мерген Меге-Баян-Далай Кара-Кес-Халбаадыр ашак Мерун-Хулук Хаан-Карацгай Улуг-Хаан Бокту-Кириш, Бора-Шээлей Анан-Даваа Олээдей-Мерген
Тыва тоолдар (О.К-Ч. Дарыманыц чыып бижээн материалдары)	С.М. Орус-оол, М. Б. Кунгаа	2014	43 44 45 46 47 48 49 50 51	Арзылац-Кара аъттыг Хунан-Кара Сарыг-Хемниц иштин чурттаан Тавын- Хаан Хан-Шилги аъттыг Хан-Кучу-Маадыр Алдын-Сарыг аъттыг Анчы-Кара Кацгай-Кара аъттыг Хайырты-Кара Анан-Даваа Олээдей-Мерген Элестей ашак Далай-Бизен-Хаан

Таблица 2

Фрагменты метаописания вариантов фольклорного текста «Бокту-Кириш, Бора-Шээлей». Программные средства формирования контекстов

Наименование текста	Кол-во слово-употреблений	Сказитель	Место рождения сказителя	Год рождения сказителя	Кем и когда записан текст	Сборник, в котором издан текст	Год и место издания сборника
Бокту-Кириш, Бора-Шээлей	15535	Салчак Чанзац	Бай-Тайгинский район	1892	1947, Тоюц Константином	«Алдай-Буучу»	1993, Тув. изд-во, Кызыл
Бокту-Кириш, Бора-Шээлей	35915	Иргит Ширинец Сундуевич	Монгун-Тайгинский район	1929	1963, Калзан Антон Каваевич	«Бокту- Кириш, Бора- Шээлей»	1995, Тув. изд-во, Кызыл
Боктуг-Кириш, Бора-Шээлей	19265	Ооржак Мацнай Намзыраевич	Сут-Хольский район	1892	1962, самозапись	«Тувинские героические сказания. Памятники фольклора народов Сибири и Дальнего Востока»	1997, Изд-во «Наука», Новосибирск
Бокту-Кириш, Бора-Шээлей	3829	Салчак Бичен Наадын-Хееевна	Бай-Тайгинский район	1901	1962, Хертек Сергей и Булугун Очур-оол (студенты КГПИ)	«Тыва туцтоолдары. Тываныцтоол чуларыныци йидугаарслед унуцматериалдары»	2012, ООО «Журналист», Абакан

При решении филологических и лингвистических задач с помощью математических методов машинного обучения и анализа формальных понятий возникает необходимость формирования бинарных контекстов на основе метаописаний текстов, хранящихся в корпусе. Для этого была разработана программа, позволяющая отбирать признаки и шкалировать качественные признаки, если таковые имеются. Программа написана на языке программирования Delphi. Отладка программы осуществлялась на компьютере с процессором Intel® Core™ i7-720QM Processor (6M Cache, 1.60 GHz) и ОЗУ размером 4 ГБ. На рисунках 1 и 2 показан пример создания контекста для задачи выявления языковых особенностей произведений тувинского героического эпоса.

Пример шкалирования признака «Ареал» представлен на рисунке 3.

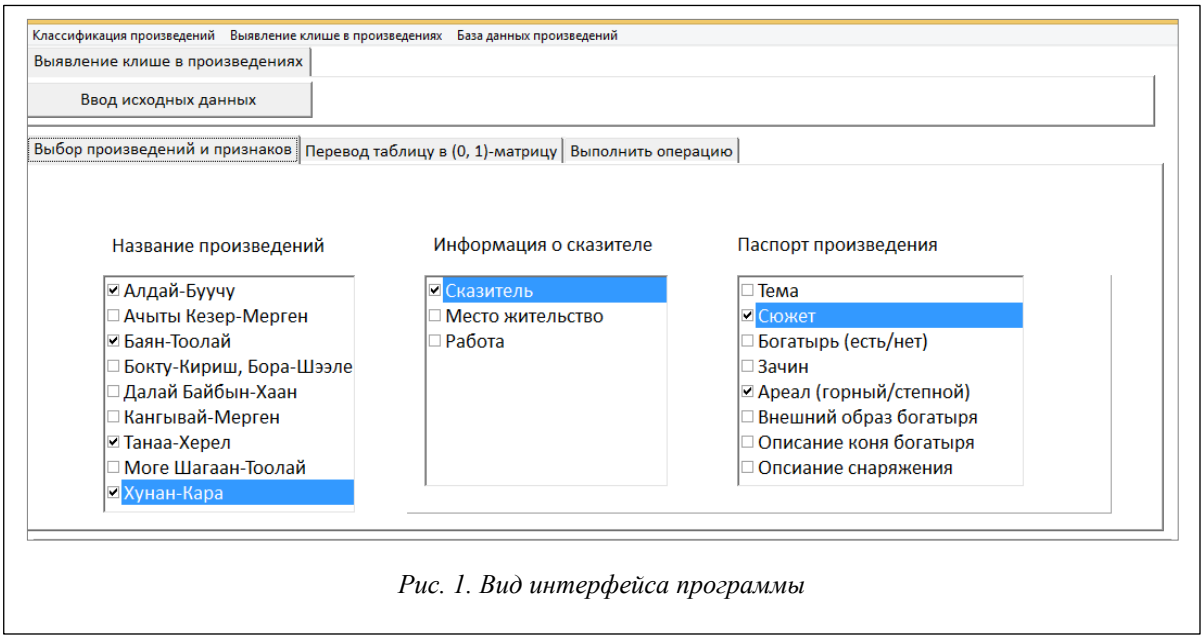

Рис. 1. Вид интерфейса программы

№	Название произведения	Сказитель	Ареал (горный/степной)	Сюжет
1	Алдай-Буучу	Кашкак Ч.М.	Горный	Сватовство
2	Баян-Тоолай	Монгуш М.С.	Степной	Сватовство
3	Танаа-Херел	Кашкак Ч.М.	Горный	Сватовство
4	Хунан-Кара	Ховалыг С.А.	Степной	Сестра добывает брату су

Рис. 2. Фрагмент контекста для произведений героического эпоса

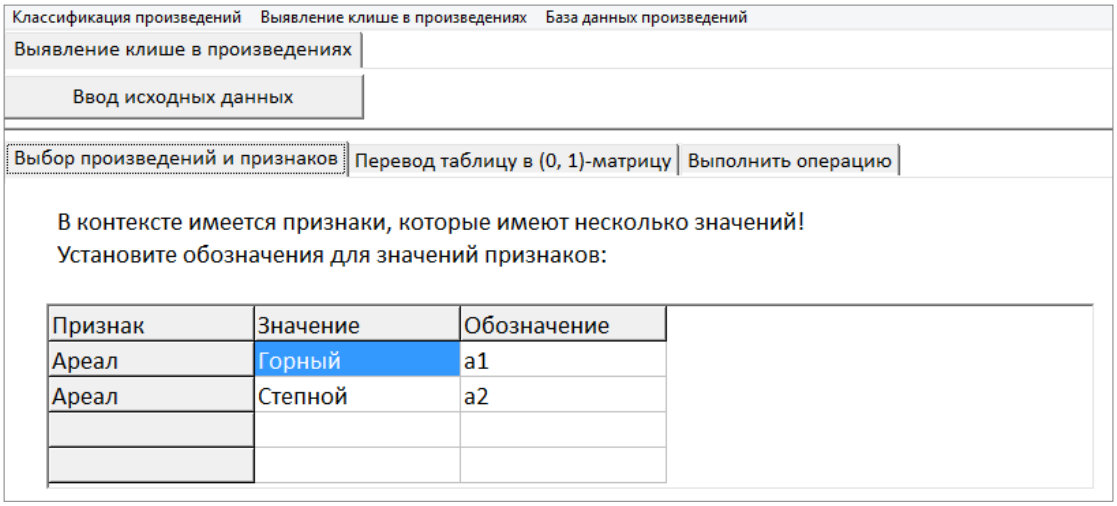

Рис. 3. Шкалирование признака «Ареал»

Заключение

Созданная БД тувинской фольклористики существенно пополняет информационную составляющую Национального корпуса тувинского языка и создает основу для расширения спектра научных исследований по изучению и сохранению национального культурного наследия тувинского народа. Работа по усовершенствованию и развитию БД тувинской фольклористики продолжается. В дальнейшем предполагается расширить эту базу не только текстами героического эпоса, но и другими жанрами фольклора. Также ведутся работы по созданию алгоритмов и программ для концептуального моделирования и бинарной классификации по прецедентам и по внедрению их в состав ПО Национального корпуса тувинского языка.

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: 14.06.2017).

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

3. Орус-оол С.М. Тувинские героические сказания (текстология, поэтика, стиль). М.: Макс Пресс, 2001. 422 с.

4. Быкова В.В., Монгуш Ч.М. Методы анализа формальных понятий в исследовании текстов тувинского фольклора // Информационные технологии и математическое моделирование (ИТММ–2016): мат. XV Междунар. конф. имени А.Ф. Терпугова «Информационные технологии и математическое моделирование» (ИТММ–2016), 12–16 сентября 2016 г., Анжеро-Судженск, Россия. Томск: Изд-во Том. ун-та, 2016. Ч. 2. С. 153–158.

5. Батура Т.В. Математическая лингвистика и автоматическая обработка текстов. Новосибирск: Изд-во НГУ, 2016. 166 с.

6. Ондар М.В. База данных текстов тувинского героического эпоса: первый этап // Новые исследования Тувы. 2016. № 4. URL: http://nit.tuva.asia/nit/article/view/616 (дата обращения: 14.06.2017).

7. Монгуш Ч.М. Анализ слабоструктурированных текстов на тувинском языке // Актуальные проблемы исследования этноэкологических и этнокультурных традиций народов Саяно-Алтая. 2015. № 3. С. 86–87.

8. Монгуш Ч.М. Задачи и методы анализа текстов тувинской словесности // Становление и развитие физико-математического образования и науки в Республике Тыва. 2014. № 1. С. 116–119.

9. Бавуу-Сюрюн М.В., Далаа С.М., Монгуш Ч.М., Ондар М.В. Тувинские героические сказания. Свид. о гос. регистр. базы данных № 2017620090, Рос. Федерация; заявл. 03.10.2016. Опубл. 19.01.2017.