УДК 81’33+821.584.6
ББК 83.3(0)9+81.23

В.В. Куканова, Е.В. Бембеев, Д.Н. Музраева

К ВОПРОСУ О КОДИРОВКЕ UNICODE ГРАФИЧЕСКОЙ СИСТЕМЫ «ТОДО БИЧИГ» И СОЗДАНИИ БАЗЫ ДАННЫХ ТЕКСТОВ НА СТАРОКАЛМЫЦКОМ ЯЗЫКЕ

* Исследование выполнено при финансовой поддержке гранта РГНФ «Национальный корпус калмыцкого языка» № 12-04-12047 (2012–2014).

Аннотация. В статье рассматриваются вопросы UNICODE-кодировки символов графической системы «тодо бичиг» («ясного письма»), описываются правила транслитерации письменных источников, база данных текстов на старокалмыцком языке.

Ключевые слова: «тодо бичиг» («ясное письмо»), ойратский язык, транслитерация, база данных, каталогизация, рукописи, печатные издания.

V.V. Kukanova, E.V. Bembeev, D. N. Muzraeva

ON THE QUESTION OF UNICODE ENCODING OF “TODO SCRIPT” GRAPHICS SYSTEM AND CREATING OF A DATABASE OF TEXTS ON OLD-KALMYK LANGUAGE

Abstract. The article examines the character of UNICODE encoding of the “Todo script” graphics system, describes the rules of transliteration of written sources, as well as the database of texts on the Old-Kalmyk language.

Keywords: *“Todo script”, Oirat language, transliteration, database, cataloging, manuscripts, blockprints.*

В последние десятилетия в филологической науке предпринимаются попытки создания синхронных и диахронных информационно-аналитических систем, на основе которых проводится описание и исследование духовного письменного наследия. Такие системы стали возможными только с применением компьютерных технологий. Представление, или публикация, созданных в Интернет ресурсов является необходимым шагом в демонстрации взаимосвязей той или иной культуры с мировым пространством. В результате письменные источники получают большую степень популяризации и ротации в научных кругах. Если говорить об оцифровке исторических документов, рукописей, памятников и сочинений религиозного характера, то возможности их использования выходят за пределы науки: они становятся доступными для более широкой аудитории.

«Ранние» тексты на старокалмыцком языке[1] пока еще не были объектом специального комплексного изучения с привлечением компьютерных технологий. Традиционно письменные источники по старокалмыцкому языку, куда можно отнести религиозные сочинения, переводную литературу, литературные памятники, деловые документы и т. д., изучались с дескриптивной точки зрения. Однако на сегодняшний день число памятников на старокалмыцком языке, введенных в научный оборот, значительно превы-

шает количество тех, которые еще не известны научному сообществу и, следовательно, не изучены с лингвистической, текстологической и иной точек зрения. Следует заметить, что, во-первых, такие тексты настоятельно нуждаются в оцифровке, поскольку срок их хранения не большой, во-вторых, «значительный по объему и ценности пласт письменного наследия калмыков и их исторических предков – ойратов – безвозвратно утрачен в период войн, в условиях политики атеизации страны и борьбы с религиозными пережитками в XX в., в годы депортации калмыков» [3, с. 9], а это значит, что памятников на старокалмыцком языке немного и то, что сейчас мы можем обнаружить, это лишь крупицы из письменного наследия калмыцкого этноса. Это еще раз подчеркивает актуальность работы по оцифровке и создания базы данных письменного наследия калмыцкого этноса. Такая работа является непременным условием дальнейшего расширения и углубления наших знаний об истории калмыцкого языка.

Создание некоего Сводного каталога и базы данных письменных источников на «тодо бичиг» является важным шагом к их изучению. Работа по оцифровке безусловно ведется[2], но следует признать, что в ней отсутствуют единые методики по сохранению оцифрованного материала, но каждый исследователь, каждый институт идут своим собственным путем в решении этого вопроса. К сожалению, отсутствие общих принципов в сохранении письменного наследия ведет к появлению разных типов каталогов, различных систем транслитерирования текстов на «тодо бичиг». Это в конечном итоге приводит к тому, что результаты этой трудоемкой работы иногда и невозможно объединить в одно целое, например, в одну универсальную базу данных, с которой могли бы работать не только ученые, но и все желающие.

Серьезным недостатком традиционного типа описаний ойратских текстов и документов является также неполное или ограниченное описание письменных источников, раскрытие содержания документа, что принципиально не изменяет способа доступа к информации. По-прежнему исследователь в поисках нужной информации должен просматривать значительное количество источников, последовательно «листая» их.

В связи с вышесказанным важнейшей задачей сегодня является решение проблемы поиска информации в созданных электронных хранилищах документов по их содержанию. Хранилища современных документов – это сами документы в текстовом формате, и их неотъемлемая часть – автоматически (автоматизированно) полученные поисковые образы. Такие информационно-поисковые возможности для хранилищ ранних рукописных и ксилографических текстов на ойратском (или старокалмыцком) письме в настоящее время отсутствуют. Реализация их представляет собой актуальную научно-практическую задачу.

Сама же графическая система, на которой писались эти тексты, – «тодо бичиг» (‘ясное письмо’) – не получила еще должной компьютерной обработки в силу ряда особенностей. Если символы монгольского письма («худам бичиг») уже получили кодировку UNICODE и уже создан пакет программ с использованием этого письма (Windows, Microsoft Office), то для «тодо бичиг» отсутствует кодировка ряда символов в соответствии с вышеупомянутыми стандартами (см. ниже таблицу соотношений графемы, глифа и кодировки Unicode, там, где отсутствует кодировка, стоит знак вопроса).

К тому же еще не разработан принцип вертикального письма как для «тодо бичиг», так и для «худам бичиг» (хотя Интернет-ресурсы предоставляют одну подобную программу, в которой соблюдается этот принцип (http://www.dusal.net/downloads/vertNote.rar), но эта программа не интегрирована в систему Microsoft Office).

Изолированная позиция	Unicode	Название буквы	Инициальная позиция		Медиальная позиция		Финальная позиция		Комментарий
			Начертание	Unicode	Начертание	Unicode	Начертание	Unicode	
ᠠ	1820	MONGOLIAN LETTER A	ᠠ	?	ᠠ	?	ᠠ ᠠ	?	В конце слова буква «А» имеет два начертания: 1) после согласной буквы «Б» «хвостик» смотрит влево; 2) после остальных согласных букв в конечной позиции «хвостик» смотрит вправо. Долгота обозначается специальным знаком долготы «ᠢ ».
ᠨ	1828	MONGOLIAN LETTER NA	ᠨ	1828	ᠨ	?	ᠨ	?	Употребляется во всех позициях. В абсолютном конце слова имеет такое же начертание, как и буква «А» («хвостик» смотрит вправо).
ᡮ	184D	MONGOLIAN LETTER TODO QA	ᡮ	184D	ᡮ	?	–	–	Употребляется в твердорядных словах в начале и середине слова.
ᡯ	184E	MONGOLIAN LETTER TODO GA	ᡯ	184E	ᡯ	?	–	–	Употребляется в твердорядных словах в начале и середине слова.
ᡳ	1845	MONGOLIAN LETTER TODO I	ᡳ	?	ᡳ	?	ᡳ ᡳ ᡳ	?	Конечная буква «И» имеет свое начертание после согласных букв «Г» и «К».
ᡵ		MONGOLIAN LETTER LA	ᡵ	182F	ᡵ	?	ᡵ	?	Употребляется во всех позициях слова.
ᡶ	182E	MONGOLIAN LETTER TODO MA	ᡶ	182E	ᡶ	?	ᡶ	184F	Употребляется во всех позициях слова.

ᡆ	1846	MONGOLIAN LETTER TODO O	ᡆ	1846	ᡆ	?	ᡆ	?	Буква обозначает твердую гласную «О». Долгая «О» обозначается знаком долготы «ᡃ ». Употребляется во всех позициях слова.
ᡇ	1847	MONGOLIAN LETTER TODO U	ᡇ	1847	ᡇ	?	ᡇ	?	Буква обозначает твердую гласную «У». Долгая «У» обозначается удвоенным написанием буквы «УУ». Употребляется во всех позициях слова.
ᡉ	1849	MONGOLIAN LETTER TODO UE	ᡉ	1849	ᡉ	?	ᡉ	?	Буква обозначает твердую гласную «Ү». Долгая «Ү» обозначается удвоенным написанием буквы «ҮҮ». Употребляется во всех позициях слова.
ᠰ	1830	MONGOLIAN LETTER SA	ᠰ	1830	ᠰ	1830	ᠰ	1830	Употребляется во всех позициях слова.
ᠱ	1831	MONGOLIAN LETTER TODO SHA	ᠱ	1831	ᠱ	1830	ᠱ	1830	Употребляется во всех позициях слова.
ᡑ	1851	MONGOLIAN LETTER TODO DA	ᡑ	1851	ᡑ	1851	ᡑ	□?	Употребляется во всех позициях слова.
ᡐ	1850	MONGOLIAN LETTER TODO TA	ᡐ	1850	ᡐ	1850	–	–	Употребляется в начале и середине слов.
ᡄ	1844	MONGOLIAN LETTER TODO E	ᡄ	1844	ᡄ	?	ᡄ	?	Обозначает мягкий гласный «Е». Долгота обозначается знаком долготы «ᡃ »
ᡗ	1855	MONGOLIAN LETTER TODO YA	ᡗ	1855	ᡗ	1855	–	–	Употребляется в начале и середине слова.

ᠷ	1837	MONGOLIAN LETTER TODO RA	ᠷ	1837	ᠷ	1837	ᠷ	?	Употребляется во всех позициях слова
ᡳᡳ	1848	MONGOLIAN LETTER TODO OE	ᡳᡳ	1848	ᡳᡳ	?	ᡳᡳ	?	Буква обозначает мягкую гласную «Ѳ». Долгая «Ѳ» обозначается знаком долготы «ᡳ ». Употребляется во всех позициях слова.
ᡳᡳ ᡳᡳ	□?	?		–	?	?	?	?	Буква обозначает согласную «Г» употребляемую в мягкорядных словах. Употребляется во всех позициях слова.
ᡳ ᡳ	1858	MONGOLIAN LETTER TODO GAA	ᡳ	183A	?	?	–	–	Буква обозначает согласную «Ка» употребляемую в твердорядных словах. Употребляется в начале и середине слова.
ᡳ	183B	MONGOLIAN LETTER TODO KA	ᡳ	183B	?	?	–	–	Буква обозначает согласную «Ке» употребляемую в мягкорядных словах. Употребляется в начале и середине слова
ᡳ	182A	MONGOLIA N LETTER TODO BA	ᡳ	182A	ᡳ	182A	ᡳ	184B	Буква обозначает согласную «Б». Употребляется во всех позициях слова
ᡳ	184C	MONGOLIAN LETTER TODO PA	ᡳ	184C	ᡳ	184C	ᡳ	184C	Буква обозначает согласную «П» употребляемую в заимствованных словах.
ᡳ	1854	MONGOLIAN LETTER TODO TSA	ᡳ	1854	ᡳ	1854	–	–	Буква обозначает согласную «Ц». Встречается в начале и середине слова. Если после этой буквы следует гласный «И», то он читается как «Ч».
ᡳ	1853	MONGOLIAN LETTER TODO JA	ᡳ	1853	ᡳ	1853	–	–	Буква обозначает согласную «Ж». Позднее нововведение. Встречается в начале и середине слова.

ᠴ	1852	MONGOLIAN LETTER TODO CHA	ᠴ	1852	ᠴ	1852	–	–	Буква обозначает согласную «Ч». Позднее нововведение. Встречается в начале и середине слова.
ᠵ	185A	MONGOLIAN LETTER TODO JIA	ᠵ	185A	?	?	–	–	Буква обозначает согласную «Ж». Позднее нововведение. Встречается в начале и середине слова.
ᠸ	1856	MONGOLIAN LETTER TODO WA	ᠸ	1856	ᠸ	1856	–	–	Буква обозначает согласную «В». Употребляется в заимствованных словах. Встречается в начале и середине слова.
ᠩ	184A	MONGOLIAN LETTER TODO ANG	–	–	ᠩ	?	ᠩ	184A	Буква обозначает согласную «Ӈ». Встречается в середине и конце слова.
ᠺ	1859	MONGOLIAN LETTER TODO HAA							Буква обозначает согласную «ХА». Употребляется в заимствованных словах. Встречается в начале и середине слова.
		L E T T							
ᠽ	185C	MONGOLIAN LETTER TODO DZA							Буква обозначает согласную «ДЗА». Употребляется в заимствованных словах. Встречается в начале и середине слова.

Другой проблемой является фонетический принцип написания слов, т. е. слова фиксировались в ранних памятниках так, как произносились. С одной стороны, это отражает ту речь, которая бытовала в обществе (в особенности это важно для лингвистов), для программистов же этот аспект является проблемным в процессе разработки распознающей программы на словарной основе, поскольку порождает большое количество вариантов написания того или иного слова. Кроме того, отдельную проблему составляет нечеткое начертание отдельных графем, графических знаков и «диакритики», что затрудняет их интерпретацию, адекватную передачу графических особенностей памятника (например, необычной лигатуры), отсутствие унификации и др.

В русле данных исследований в Калмыцком институте гуманитарных исследований РАН ведется работа по созданию Национального корпуса калмыцкого языка, одним из направлений которого является разработка подкорпуса «ранних» текстов. В рамках проекта проводится анализ рукописных и печатных источников XVII–XIX вв., выяв-

ляются их палеографические и лексические характеристики, среди которых можно перечислить особенности почерка переписчика или шрифта текста, формат рукописи, качество бумаги, чернил и т. п. следы времени. Предпринятая работа в дальнейшем существенно облегчит работу по вводу и обработке калмыцких и ойратских текстов и изображений, систем оптического распознавания, систем информационного поиска и автоматического индексирования документов.

В 2012 г. был проведен эксперимент в целях обнаружения тех или иных проблем при автоматической обработке текстов на материале фототипического издания текста, который опубликовал профессор Санкт-Петербургского университета А. М. Позднеев в 1897 г. под названием «Сказание о хождении в Тибетскую страну малодербетовского Бааза-бакши»[3] [5]. В ходе пилотного анализа «раннего» текста был выявлен ряд проблем, касающихся транслитерации текста «тодо бичиг», орфографии текста, омонимии словоформ, разметки текста, использования диакритических знаков и т. д. [1]. Одной из таких проблем явилось то, что знак «:», обозначающий долготу гласного и по традиции используемый в латинской транслитерации текстов ойратских текстов, программы ошибочно распознавали как делиметр, т. е. разделитель (так же, как дефис или пробел),. В результате данный знак «:» в обрабатываемых текстах пришлось заменить на знак «¯». Эти и ряд других проблем были учтены при дальнейшей обработке массива текстов на «тодо бичиг», и, как результат, был выработан алгоритм обработки текстов на «тодо бичиг». . Ниже приведен список правил транслитерации, при этом опирались на следующие работы: [6; 2].

ПРАВИЛА ТРАНСЛИТЕРАЦИИ

1. Буквы «тодо бичиг» традиционно транслитерируются латиницей, однако было решено ряд графем и глифов транслитерировать особыми символами для упрощения автоматического анализа текстов. Ниже приведена таблица графем.

№	Графема тодо бичиг	Символ, используемый при транслитерации		
1.	ᠠ	a	a долгая	ā
2.	ᠡ	e	e долгая	ē
3.	ᠢ	i	i долгая	ī
4.	ᠣ	o	o долгая	ō
5.	ᠥ	ö	ö долгая	ȫ
6.	ᠤ	u	u долгая	ū
7.	ᠦ	ü	ü долгая	ǖ
8.	ᠨ	n		
9.	ᠬ	x		
10.	ᠭ	γ		
11.	ᠪ	b		
12.	ᠰ	s		
13.	ᠱ	š		
14.	ᠲ	t		
15.	ᠳ	d		

№	Графема тодо бичиг	Символ, используемый при транслитерации				
16.	[illegible]	l				
17.	[illegible]	м				
18.	[illegible]	c	перед гласной i	č	[illegible]	c̓
19.	[illegible]	z	перед гласной i	ǰ	[illegible]	ž
20.	[illegible]	y				
21.	[illegible]	g				
22.	[illegible]	ḳ				
23.	[illegible]	k				
24.	[illegible]	q				
25.	[illegible]	r				
26.	[illegible]	v				
27.	[illegible]	ŋ				
28.	[illegible]	ha				
29.	[illegible]	p				
30.	[illegible]	f				

2. Конец предложения (•••) маркируется знаком (=). Запятая обозначается (,).

3. Конец абзаца (или текста ❖) маркируется (==).

4. Падежные окончания, которые были написаны отдельно от слова (через пробел), транслитерируются через дефис (-). – morin-du

5. Начало листа обозначается квадратными скобками, внутри которых помещают номер листа. Например: [1a] или [1b].

6. Номер строки обозначается в круглых скобках (). Строки пишутся через знак абзаца (через Enter), т. е. с новой строки.

7. Все собственные имена следует писать с заглавной буквы для облегчения дальнейшего компьютерного анализа текста.

8. Предложения не следует начинать с заглавной буквы.

9. Если границы строки проходят внутри слова, то мы маркируем этот факт косой чертой (/). При совпадении границы строки и написания аффикса отдельно от слова на другой строки, используется комбинация символов: (-/).

10. Если в тексте имеется вставка слога, слова или предложения и если она сделана тем же самым почерком, то мы обозначаем такие вставки в фигурных скобках {}.

11. Если в тексте имеется вставка слога, слова или предложения и если она сделана другим почерком, то мы обозначаем такие вставки в фигурных и угольных скобках {<>}.

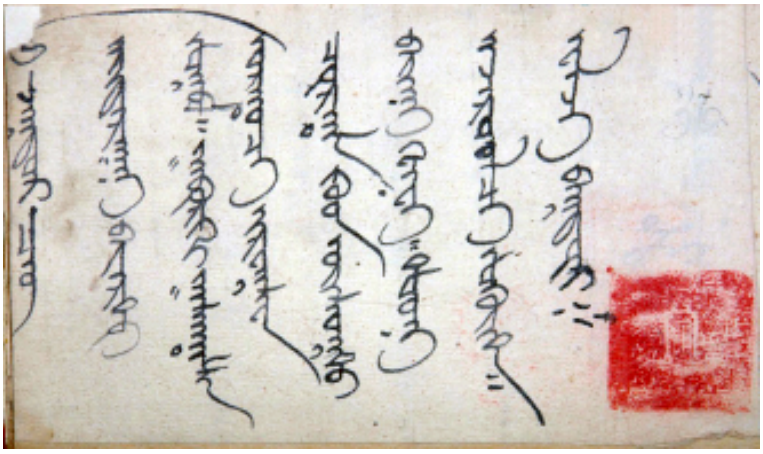
12. Если в тексте имеется неразборчиво написанный отрезок текста, то такого рода отрывки мы помечаем <...>.

13. Неуверенность исследователя при транслитерации отмечается знаком вопроса в круглых скобках (?).

14. Комментарий исследователя помещается в квадратных скобках. Например, [текст утрачен].

15. Неязыковые компоненты обозначаются двойным знаком (*). например, *квадратная печать на квадратной письменности, выполненная красной тушью*.

Приведем пример транслитерированного текста.



[4: (Из писем хана Аюки и Чагдаржапа. 1714 г.)]

[1a]
(1) Cagdor ǰab
(2) Ayidarxani bayartu
(3) öqbö= xaburki xasaγāsa
(4) orγoǰi iregsen
(5) zurγān kū̃n üyisüqtü
(6) bayinai genei= töüni
(7) acaroulǰi ögüyita=
(8) elči Baqdor==
квадратная печать на квадратной письменности, выполненная красной тушью

Кроме правил транслитерации, был разработан электронный ресурс по каталогизации и транслитерации текстов на старокалмыцком языке (прогр. А.Ю. Каджиев). Портал находится по адресу http://kalmcorpora.ru/todo. На данный момент он открыт только для исполнителей проекта. База данных по метаописанию спроектирована в mySQL, создан web-ориентированный программный интерфейс для транслитерации оцифрованных текстов на старокалмыцком языке. Пользователь дает библиографическое описание документа, который загружается на сервер по следующим атрибутам:

1) заголовок;
2) название, данное исследователем;
3) заголовок по титульному листу;
4) заголовок по первой строке документа;
5) заголовок по колофону;
5) маргинальное название;
6) автор (если имеется);
7) переводчик (если имеется);
8) переписчик (если имеется);
9) тема (указать кратко);
10) описание;

11) источник;
12) формат листа;
13) формат рамки (если имеется);
14) цвет чернил;
15) печать (если имеется);
16) тип письма (уставной, скоропись);
17) оценка почерка (разборчиво, неразборчиво, частично разборчиво);
18) оценка качества («5» – отличное качество; «4» – хорошее качество; «3» – удовлетворительное качество; «2» – плохое качество).

На данном этапе разработки проекта на указанный сайт загружено 100 архивных документов, из них транслитерировано 88 листов (около 5 000 токенов).

Таким образом, задача сохранения духовного наследия наших предков для будущих поколений, которая стоит перед исследователями современности, носит ретроспективный характер и охватывает самый широкий круг вопросов – от текстологии и диалектологии до сравнительно-исторического изучения словоформ, словосочетаний и т. д. Эта работа может привести в дальнейшем к реконструкции ойратских и общемонгольских древностей на вербальном уровне.

Примечания

[1] Здесь имеется в виду тексты, написанные на «тодо бичиг» (‘ясном письме’). Даннаяграфическая система, напомним, была создана Зая-пандитой в 1648 г. Ойратское письмо, как и старомонгольское, имеет вертикальное направление, буквы в слове и слова пишутся сверху вниз. Слова в столбцах разделяются пробелами, столбцы располагаются слева направо. Большинство букв имеет три различных написания – в начале, середине и конце. Более того, для обозначения звуков, отсутствующих в ойратском (старокалмыцком) языке, используются дополнительные буквы-«галики». Они встречаются в основном в религиозных текстах для обозначения заимствований из тибетского или русского языка, санскрита. Орфография «тодо бичиг» в основном фонетическая, т. е. каждая буква отражает один звук, что является главным отличием от полифонного старомонгольского письма.

[2] На современном этапе исследователями основное внимание уделяется задачам поиска, создания каталогов и сохранения исторических памятников. Основным методом переноса на новые носители является оцифровка данных, подразумевающая факсимильное копирование источников и сопровождение их библиографическими и археографическими данными. К примеру, такая работа проводится Общественной организацией «Тод номын гэрэл» (Монголия), которая совместно с Американским центром монголоведения (The American Center for Mongolian Studies – ACMS) разместили на сайте 140 рукописных текстов на «ясном письме» (http://www.dlir.org/archive/orc-exhibit/ items/browse/collection/7).

[3] Рукопись была приобретена у автора Бааза Менкеджуева профессором А.М. Позднеевым, который позднее опубликовал ее с переводом и комментариями. Оригинал рукописи до сих пор не обнаружен. Издание было посвящено XI международному съезду ориенталистов в Париже. Сочинение состоит из 278 страниц: предисловие – 18 страниц (пагинация римскими цифрами, постраничная); перевод занимает 130 страниц (пагинация арабскими цифрами, общая, постраничная); текст на «Тодо бичиг» – 120 страниц (пагинация арабскими цифрами, общая, постраничная. На странице 12 строк, сверху вниз, слева направо). Материалом для нашего исследования послужило данное фототипическое издание текста на старокалмыцкой письменности «Тодо бичиг».

Список литературы

1. Бембеев Е.В. Опыт квантитативной обработки текста на старокалмыцком языке: количественные характеристики // Вестник Калмыцкого института гуманитарных исследований РАН. 2012. № 2. С. 163–168.

2. Музраева Д.Н. Опыт археографического описания и текстологического анализа рукописного перевода Тугмюд-гавджи (на материале VI главы Oülgurun dalai «Моря притч») // Вестник Калмыцкого интститута гуманитарных исследований РАН. 2012. № 3. С. 167–185.

3. Музраева Д.Н. Буддийские письменные источники на тибетском и ойратском языках в коллекциях Калмыкии. Элиста: ЗАОр «НПП „Джангар"», 2012. 224 с.

4. Национальный архив Республики Калмыкия (НА РК). Ф. 36. Оп. 1. Д. 2. Л. 56.

5. Сказание о хождении в тибетскую страну малодербетовского Бааза-бакши / пер. и коммент. А.М. Позднеева. СПб., 1897. 18+130+120 с.

6. Яхонтова Н.С. Ойратский литературный язык XVII в. М.: Вост лит., 1996. 152 с.