***Е.В. Бембеев,***
*КИГИ РАН*

# ОПЫТ КВАНТИТАВНОЙ ОБРАБОТКИ «РАННЕГО» ТЕКСТА НА КАЛМЫЦКОМ ЯЗЫКЕ (НА ПРИМЕРЕ МАТЕРИАЛА НАЦИОНАЛЬНОГО КОРПУСА КАЛМЫЦКОГО ЯЗЫКА)*

**Статья подготовлена при поддержке проекта «Национальный корпус калмыцкого языка» подпрограммы фундаментальных исследований Президиума РАН «Создание и развитие корпусных ресурсов по языкам народов России» программы «Корпусная лингвистика» (2012–2014) и проекта РГНФ «Национальный корпус калмыцкого языка» (12-04-12047, тип «в»)*

В последние годы большой интерес приобретают исследования и экспликация языковых явлений с помощью методов количественной математики, так называемой квантитативной лингвистики. Одним из реальных оснований применения статистических методов в изучении языка и речи (текста) следует признать объективную присущность языку количественных признаков, т. е. квантитативных характеристик. Повторяемость (рекуррентность, периодичность) языковых, в том числе лексических единиц, их воспроизведение в различных текстах является наиболее важным условием квантификации языкового материала и применения методов количественной математики для его анализа [1; 284]. Между тем следует помнить, что квантитативный подход, упрощая языковую реальность, способен охватить лишь определенный аспект языка и речи, которые не всегда удается обнаружить в ходе качественного анализа.

Одной из важнейших задач квантитативной лингвистики является составление частотных словарей, необходимость использования которых для решения прикладных и исследовательских задач несомненна. Частотный словарь – это модель особым образом преобразованного текста, модель распределения частот употребления единиц в тексте. Как отмечает в своей работе В.А. Долинский, словарь подобного рода «…включает в себя упорядоченный список слов или других языковых единиц (словоформы, словосочетания), которые зарегистрированы составителем в обследованном им тексте, фрагменте текста или корпусе текстов и снабжены данными о частоте их употребления в тексте (речи). С его помощью можно попытаться ответить на вопросы: как много слов в языке (тексте), с какой интенсивностью они используются в речи, какие из них предпочтительнее в той или иной сфере коммуникации у того или иного автора и т. д. [1; 285].

В настоящей статье нами предпринята попытка квантитативного анализа «раннего» текста, обработка которого является пилотной в процессе создания Национального корпуса калмыцкого языка с целью выявления проблем в автоматической обработке массива текстов, написанных на «тодо бичиг», а затем транслитерированных на латиницу. Эксперимент был проведен на материале фототипического издания текста, который в 1897 г. под названием «Сказание о хождении в Тибетскую страну малодербетовского Бааза-бакши» опубликовал профессор Санкт-Петербургского университета А.М. Позднеев [2]. Данный памятник является единственным образцом из сохранившихся до настоящего времени письменных свидетельств оригинального жанра хождений в калмыцкой литературе. Язык текста «Сказания…» неразрывно связан с личностью автора, временем, местом и условиями, в которых он жил .

Прежде чем приступить к обработке данных, необходимо отметить, что анализу подвергаются не лексемы, а словоформы, обладающие реальной частотностью в языке текста. Интересно также отметить, что составители частотных словарей отмечают привычность основной словарной единицы – лексемы, а также тот факт, что при сведении словоформ в лексемы лингвисты могут использовать разные принципы, что приводит к более высокой доле субъективности в количественных показателях по лексемам по сравнению с количественными данными по словоформам. Между тем по оценке Л. Лённгрена: «…количественные языковые факты, опирающиеся только на уровень словоформ, являются более объективными и надёжными» [3; 28–29]. Таким образом, в качестве единицы описания выступает слово «от пробела до пробела» в той грамматической форме, в которой она употреблена.

Текст «Сказания…» обрабатывался в специальных лингвистических программах TLEX Corpus 5.1 и Language Explorer. В ходе обработки текста возникло несколько проблем. Например, знак долготы гласных, по традиции это двоеточие, которое ставится после гласной буквы, пришлось заменить на «¯», поскольку указанные программы ошибочно распознавали его как делиметр, т. е. разделитель (так же, как дефис или пробел).

Общий список словоформ по частоте представлен более 5 000 единицами, включая имена собственные. Каждой словоформе приписан ранг, а также указана абсолютная частота по всему тексту в целом. Ниже приведен список наиболее самых частотных словоформ, которые были употреблены в тексте более 50 раз. Таблица организована по принципу убывания общей частоты встречаемости словоформ. Дополнительными графами в этом разделе являются «относительная частица» и «часть речи». В последней приводятся название частей речи в соответствии с общепринятыми пометами, используемыми в корпусе.

*Список наиболее частотных словоформ*
*в «Сказании о хождении в Тибетскую страну малодербетовского Бааза-бакши»*

| Ранг | Словоформа | Частота | Относительная частота (%) | Часть речи |
|---|---|---|---|---|
| 1. | ene | 482 | 2,16 | PRON |
| 2. | nige | 288 | 1,29 | NUM |
| 3. | tere | 233 | 1,04 | PRON |
| 4. | bide | 227 | 1,02 | PRON |
| 5. | ügei | 212 | 0,95 | PART |
| 6. | geži | 188 | 0,84 | V |
| 7. | biden | 185 | 0,83 | PRON |
| 8. | basa | 176 | 0,79 | CONJ |
| 9. | bayinai | 170 | 0,76 | V |
| 10. | küün | 168 | 0,75 | N |
| 11. | tegēd | 165 | 0,74 | CONJ |
| 12. | gene | 147 | 0,66 | V |
| 13. | gedeg | 143 | 0,64 | V |
| 14. | ulus | 139 | 0,62 | N |
| 15. | yuuman | 132 | 0,59 | N |
| 16. | qoyor | 129 | 0,58 | NUM |
| 17. | yeke | 123 | 0,55 | ADJ |
| 18. | bolōd | 121 | 0,54 | V |
| 19. | čigi | 109 | 0,49 | PART |
| 20. | yabugsan | 104 | 0,47 | V |
| 21. | bayidag | 98 | 0,44 | V |
| 22. | yabād | 92 | 0,41 | V |
| 23. | düngge | 91 | 0,41 | POST |
| 24. | ödör | 89 | 0,40 | N |
| 25. | bolon | 81 | 0,36 | CONJ |
| 26. | yabuži | 79 | 0,35 | V |
| 27. | cagtu | 76 | 0,34 | N |
| 28. | qonogson | 76 | 0,34 | V |
| 29. | γazar | 76 | 0,34 | N |
| 30. | γurbun | 75 | 0,34 | NUM |
| 31. | irebe | 74 | 0,33 | V |
| 32. | kürtele | 73 | 0,33 | POST |
| 33. | cai | 72 | 0,32 | N |
| 34. | kiyid | 72 | 0,32 | N |
| 35. | žige | 72 | 0,32 | PART |
| 36. | γazartu | 71 | 0,32 | N |
| 37. | gegen | 69 | 0,31 | N |
| 38. | blama | 66 | 0,30 | N |
| 39. | bayigsan | 65 | 0,29 | V |
| 40. | dēre | 63 | 0,28 | ADV |
| 41. | bi | 62 | 0,28 | PRON |
| 42. | duunai | 62 | 0,28 | N |
| 43. | gēd | 61 | 0,27 | V |
| 44. | sayin | 61 | 0,27 | ADJ |
| 45. | žigen | 57 | 0,26 | PART |
| 46. | ireži | 56 | 0,25 | V |
| 47. | yabuqu | 51 | 0,23 | V |

Анализ представленной выборки с частотой 50 и выше показывает, что наиболее употребительными словоформами является группа указательных и личных местоимений: *ene* ‘этот, эта, это’ (482), *tere* ‘тот, та, то’ (233), *bide* ‘мы’ (227), *biden* ‘мы’ (185), *bi* ‘я’ (62). Среди глагольных форм наиболее употребительными являются: *geži* ‘говорил’ (188), *bayinai* ‘есть, быть’ (170) *gene* ‘говорит’ (147), *gedeg* ‘говоря’ (143). Среди

наиболее употребительных словоформ большой процент занимают глаголы с семантикой движения, что обуславливается характером памятника, описывающего путешествие в далекую страну. Например: *yabugsan* 'ушедший' (104), *yabād* 'идя' (92), *yabuži* 'ушел' (79), *irebe* 'пришел' (74), *ireži* 'пришел' (56), *yabuqu* 'уйдет' (51). Существительные занимают свою частотную позицию, начиная с 10 ранга: *küün* 'человек' (168), *ulus* 'народ, люди' (139), *yuuman* '(свои) вещи' (132).

Принадлежность автора к религиозной деятельности, а также цель хождения – поклонение святым и буддийским реликвиям также находят отражение в частоте употребления «буддийской» лексики: *kiyid* 'монастырь' (72), *gegen* 'гегян, светлость' (69), *blama* 'лама, учитель' (66).

Кроме того, частотный словарь «ранних» текстов позволяет проводить типологические исследования между разными подъязыками (авторскими стилями) калмыцкого языка. Путем различной сортировки и группировки данных из одной и той же словарной базы можно получить алфавитный список словоформ с указанием частоты, порядок убывания частоты, общий список словоформ по частоте различные словари текста и другие статистические сведения.

Таким образом, создание подкорпуса «ранних текстов» в рамках Национального корпуса калмыцкого языка является важной и актуальной задачей. Изучение ранних текстов носит ретроспективный характер и охватывает самый широкий круг вопросов – от текстологии и диалектологии до сравнительно-исторического изучения словоформ, словосочетаний и т. д., что часто приводит в свою очередь к реконструкции ойратских и общемонгольских древностей на вербальном уровне. Нередко в ранних текстах фиксируются лексемы и целые последовательности лексем, которые не встречаются в современных данных языка. Вместе с тем в ходе анализа «раннего» текста выявлен ряд проблем, касающихся транслитерации текста с «тодо бичиг», орфографии текста, омонимии словоформ, разметки текста, использования диакритических знаков и т. д. Эти проблемы буду учтены впоследствии при обработке массива текстов на «тодо бичиг».

**Литература**

1. Долинский В. А. Квантитативная лингвистика в исследовании текста // Алфавит: Строение повествовательного текста. Синтагматика. Парадигматика. – Смоленск: СГПУ, 2004. С. 283–324.

2. Сказание о хождении в тибетскую страну малодербетовского Бааза-бакши / пер. и коммент. А. М. Позднеева. – СПб., 1897.

3. Лёнгрен Л. (ред.). Частотный словарь современного русского языка. – Uppsala, 1993.

***

***Т.Н. Богрданова,***
*Калмыцкий государственный университет*

## ОЙРАТСКО-КАЛМЫЦКАЯ СКАЗКА НА АНГЛИЙСКОМ ЯЗЫКЕ: ОСОБЕННОСТИ ПЕРЕВОДЧЕСКОЙ СТРАТЕГИИ И ТЕХНИКИ

В настоящей статье остановимся на английском переводе сборника «Сидди-кюр» в плане обсуждения переводческой стратегии и некоторых приемов перевода, что представляется актуальным в отношении текстовой коммуникации так называемых дистантных культур, а также особенностей европейских переводческих традиций в освоении «экзотического» фольклорного материала [2; 55–63].

Интерес в XIX в. к фольклору монголов в Европе во многом обусловлен их примечательной ролью в распространении индийских сказаний [7, XI]. В связи с этим и были осуществлены немецкие переводы «Сидди-кюра», наиболее популярного сборника сказок в монгольском мире, в отношении которого связь с индийским прототипом («Двадцать пять рассказов Веталы») считалась установленной. Вслед за немецкими переводами монгольско-ойратские сказки были переведены на английский язык (Рэчел Баск, 1873 г.) [8]. В плане общей характеристики немецкого перевода Б. Юльга, одного из основных источников для английского варианта, заметим, что выполненный в рамках традиции перевода, сложившейся к этому времени в Германии, он строго следует особенностям оригинала [7, XVI], [6, 130, 150]. Так, сравнительный текстуальный анализ немецкого текста с известным переводом акад. Б.Я. Владимирцова, выполненных в разное время и с различных списков, показывает, что в целом значительных разночтений между параллельными текстами нет; авторы переводов следовали по возможности более верно монгольско-ойратским письменным оригиналам, которые в данном случае, по-видимому, практически совпали. Оба перевода представляют образец научного перевода фольклорного материала, что очевидно, в частности, при их сопоставительном анализе с английским вариантом, на котором остановимся далее более подробно.

Подчеркивая сложность транспонирования столь «экзотического» материала для англоязычного читателя, несмотря на посредство другого европейского языка, в предисловии автор перевода отмечает необходимость редакции переводимых текстов. По мнению Р. Баск, калмыцкие рассказы большей частью «непривычны» для «тех, кто воспитан в христианском духе и современной культуре» и отражают мышление столь же отличное от европейского читателя, сколь далек и язык, на котором они написаны [8, vi-vii]. Эти предварительные замечания переводчика представляют, на мой взгляд, особый интерес, поскольку отражают философию перевода, весьма характерную для английской переводческой традиции в передаче фольклорных