

УДК 81'33 + 811.11-112 + 811.512.37  
ББК 81.23(2Рос=Калм)

## ОПЫТ КВАНТИТАТИВНОЙ ОБРАБОТКИ ТЕКСТА НА СТАРОКАЛМЫЦКОМ ЯЗЫКЕ: КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ\*

*Е. В. Бембеев*

С развитием информационных технологий в языкознании широкое распространение приобретают методы математического анализа. В современном монголоведении квантитативный подход совсем недавно стал использоваться при исследовании языковых реалий, поскольку только стали создаваться лингвистически аннотированные корпуса монгольских языков [Бадагаров 2008; Бадмаева и др. 2008; Крылов 2012; Куканова 2011а; 2011б; Ринчинов 2009].

Повторяемость языковых, в том числе лексических, единиц и их воспроизведение в различных текстах является наиболее важным условием в количественном описании языкового материала и применения математических методов в лингвистике для его анализа [Долинский 2004: 284]. Квантитативный метод позволяет количественно описывать поведение различных языковых единиц (букв, фонем, морфем, слов, конструкций и т. п.) в письменном тексте (так же как и в устном): частоту употребления тех или иных единиц, их распределение в текстах разного жанра, сочетаемость с другими единицами, выявление разного рода корреляций (например, корреляция частотности слов и изменения их звуковой формы, так называемые аллегривные формы) и т. п. «Одновременно накапливается обобщенная количественная информация о классах единиц, о языковых конструкциях (напр., данные о средней длине слова или предложения, о частоте употребления к.-л. грамматических форм в тех или иных синтаксических функциях и т. п.). Такая информация углубляет описание единиц языка» [Шайкевич 1990: 231].

Возьмем к примеру проблему категории множественности. Простого описания образования форм множественного числа существительных в калмыцком языке явно недостаточно, требуется более глубокое исследование данной категории, в частности и в типологическом аспекте, например в сопоставлении с материалом русского языка. Количественные характеристики текстового

материала тех или иных единиц поможет понять различия в типологическом плане и более полно и основательно проанализировать эту категорию как в функционально-семантическом, так и в грамматическом аспектах. Однако при этом не стоит забывать, что лингвостатистический подход несколько упрощает языковую реальность и охватывает лишь определенный пласт языка и речи.

Актуальность проблемы изучения количественных характеристик в калмыцком языке обусловлена и тем обстоятельством, что большинство этих характеристик до сих пор неизвестны ученым из-за отсутствия представительных и хотя бы относительно сбалансированных корпусов калмыцкого языка. Именно на таком материале можно будет применить дистрибутивно-статистические методы, позволяющие составлять частотные словари и квантитативные грамматики, описывающие частотность единиц лексикологии, дериватологии, морфологии и синтаксиса. Квантитативный подход позволяет классифицировать сами тексты в соответствии с языковыми стилями и жанрами, в рамках которых эти тексты создавались. Поскольку различия между этими стилями и жанрами «носят преимущественно статистический характер», то таким образом можно основать статистическую стилистику калмыцкого языка, описывающую и классифицирующую тексты на строгой объективной базе [Шайкевич 1990: 231]. Такая выборка выводит исследователя текста за рамки одного произведения или творчества одного автора и позволяет отсеять индивидуальные явления от общих, тем самым выделяя универсальное и специфическое для каждого отдельного текста или каждого отдельного автора, а также языковой системы в целом. Конечно, в идеале частотное, кон-текстологическое и концептуально-текстовое направления изучения текста должны не исключать, а гармонично дополнять друг друга. Квантитативный метод исследования предполагает составление частотных слова-

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

рей, необходимость использования которых для решения прикладных и исследовательских задач несомненна.

Частотный словарь представляет собой модель особым образом преобразованного текста, модель распределения частоты 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.» [Долинский 2004: 285]. Другими словами, речь идет о привычном типе словаря лексем с частотной характеристикой. Различают частотный словарь и ранговый словарь, под последним понимается разновидность первого, в которой инвентаризуемые единицы расположены в порядке убывания употребительности (или то же самое, но только в порядке возрастания рангов).

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 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 [Сказание... 1897]<sup>1</sup>. Данный

<sup>1</sup> 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).

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 на этом этапе подвергались не лексемы, а словоформы<sup>2</sup>, обладающие реальной частотностью в языке текста. 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. Леннгрена, «...количественные языковые факты, опирающиеся только на уровень словоформ, являются более объективными и надежными» [Леннгрен 1993: 28–29]. Таким образом, в качестве единицы описания выступает слово «от пробела до пробела» в той грамматической форме, в которой она употреблена. Поскольку работа велась по тексту с неснятой омонимией, то иногда статус словоформ получают не собственно словоформы, а «дизъюнктивные пучки (частично) омонимических словоформ» [Крылов 2012: 88].

Текст «Сказания...» обрабатывался в различных лингвистических программах. В ходе его обработки возникло несколько проблем. Например, знак «:», обозначающий долготу гласного и по традиции используемый в транслитерации текстов на латиницу, пришлось заменить на «\_», поскольку программы ошибочно распознавали его как дефис, т. е. разделитель (так же, как дефис или пробел).

Текст состоит из более 21 тыс. словоупотреблений, а 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 частотных словоформ, которые были употреблены в тексте более 50 раз (см. таблицу 1). Табли-

<sup>2</sup> В нашем случае термин *словоформа* понимается как «слово (лексема) в некоторой грамматической форме (в частном случае — в единственно имеющейся у слова форме), напр.: „сад“, „садами“, „белый“, „белую“, „пишет“, „вчера“» [Зализняк 1990].

ца организована по принципу убывания общей частоты встречаемости словоформ. Дополнительными графами в этом разделе являются «относительная частица» и «часть речи». В последней приводятся названия частей речи в соответствии с общепринятыми пометами, используемыми в Национальном

корпусе калмыцкого языка. В таблице используется знак «~», маркирующий наличие омонимичной формы. Таблица состоит из 6-ти столбцов: А — ранг, В — словоформа, С — перевод словоформы, D — частота, E — относительная частота в %, F — предполагаемая часть речи.

**Таблица 1.** Список наиболее частотных словоформ в «Сказание о хождении в Тибетскую страну малодербетовского Бааза-бакии»

| А   | В        | С                                          | D   | E    | F         |
|-----|----------|--------------------------------------------|-----|------|-----------|
| 1.  | ene      | этот, он, она, оно, они                    | 482 | 2,16 | PRON      |
| 2.  | nige     | один, раз                                  | 288 | 1,29 | NUM       |
| 3.  | tere     | тот, он, она, оно, они                     | 233 | 1,04 | PRON      |
| 4.  | bide     | мы                                         | 227 | 1,02 | PRON      |
| 5.  | ügei     | нет, не                                    | 212 | 0,95 | PART      |
| 6.  | geži     | что [изъясн. союз], сказав [дее-причастие] | 188 | 0,84 | CONJ~CONV |
| 7.  | biden    | мы                                         | 185 | 0,83 | PRON      |
| 8.  | basa     | тоже, также                                | 176 | 0,79 | CONJ      |
| 9.  | bayinai  | является, имеется                          | 170 | 0,76 | V         |
| 10. | küün     | человек                                    | 168 | 0,75 | N         |
| 11. | tegēd    | поэтому                                    | 165 | 0,74 | CONJ      |
| 12. | gene     | говорят                                    | 147 | 0,66 | V         |
| 13. | gedeg    | говоря                                     | 143 | 0,64 | CONV      |
| 14. | ulus     | люди, народ                                | 139 | 0,62 | N         |
| 15. | yuuman   | [показатель ремы], вещь                    | 132 | 0,59 | N         |
| 16. | qoyor    | два                                        | 129 | 0,58 | NUM       |
| 17. | yeke     | большой, очень                             | 123 | 0,55 | ADJ       |
| 18. | bolōd    | сделав, а, что касается; [показатель темы] | 121 | 0,54 | CONV      |
| 19. | čigi     | же                                         | 109 | 0,49 | PART      |
| 20. | yabugsan | ушел, ушедший                              | 104 | 0,47 | V~CONV    |
| 21. | bayidag  | являющийся, имеющийся                      | 98  | 0,44 | PTCPL     |
| 22. | yabād    | уходя                                      | 92  | 0,41 | CONV      |
| 23. | düngge   | подобно [послелог сравнения]               | 91  | 0,41 | POST      |
| 24. | ödör     | день                                       | 89  | 0,40 | N         |
| 25. | bolon    | и [союз]                                   | 81  | 0,36 | CONJ      |
| 26. | yabuži   | ходив                                      | 79  | 0,35 | CONV      |
| 27. | cagtu    | временем                                   | 76  | 0,34 | N         |
| 28. | qonogson | ночевал, ночующий                          | 76  | 0,34 | V~PTCPL   |
| 29. | γazar    | земля                                      | 76  | 0,34 | N         |
| 30. | γurbun   | три                                        | 75  | 0,34 | NUM       |
| 31. | irebe    | пришел                                     | 74  | 0,33 | V         |
| 32. | kürtele  | до тех пор                                 | 73  | 0,33 | POST      |
| 33. | cai      | чай                                        | 72  | 0,32 | N         |
| 34. | kiyid    | монастырь                                  | 72  | 0,32 | N         |

|     |          |                                                  |    |      |      |
|-----|----------|--------------------------------------------------|----|------|------|
| 35. | žige     | точно, правда [частица]                          | 72 | 0,32 | PART |
| 36. | γazartu  | на земле                                         | 71 | 0,32 | N    |
| 37. | gegen    | гегян, святой                                    | 69 | 0,31 | N    |
| 38. | blama    | лама                                             | 66 | 0,30 | N    |
| 39. | bayigsan | был                                              | 65 | 0,29 | V    |
| 40. | dēre     | на [последлог], наверху, сверху                  | 63 | 0,28 | ADV  |
| 41. | bi       | я                                                | 62 | 0,28 | PRON |
| 42. | duunai   | километр, [расстояние слышимости звука человека] | 62 | 0,28 | N    |
| 43. | gēd      | сказав                                           | 61 | 0,27 | CONV |
| 44. | sayin    | Хорошо, хороший                                  | 61 | 0,27 | ADJ  |
| 45. | žigen    | точно, правда [частица]                          | 57 | 0,26 | PART |
| 46. | ireži    | пришел                                           | 56 | 0,25 | V    |
| 47. | yabuqu   | ходить                                           | 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 *bayinai* 'есть, быть' (170) *gene* 'говорит' (147), *gedeg* 'говоря' (143). Существуют большое количество морфологической омонимии. Так, например, слово *geži* употребляется в тексте 188 раз, она может означать изъяснительный союз 'что' или деепричастие 'сказав'. Большой процент употребления 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), *уиутан* '(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).

В приведенной ниже таблице 2 показана частотность частей речи с неснятой омонимией, поскольку снятие морфологической омонимии в старописьменном тексте придется проводить, видимо, вручную.

**Таблица 2.** Список частотности частей речи тексте памятника «Сказание о хождении в Тибетскую страну малодербетовского Бааза-бакиш»

| №   | Часть речи | Кол-во словоформ | %        |
|-----|------------|------------------|----------|
| 1.  | N          | 9095             | 42,11818 |
| 2.  | ADJ        | 934              | 4,325276 |
| 3.  | NOM        | 18               | 0,083356 |
| 4.  | N~ADJ      | 271              | 1,254978 |
| 5.  | ADV        | 766              | 3,547282 |
| 6.  | ADV~POST   | 143              | 0,662221 |
| 7.  | NUM        | 1195             | 5,533945 |
| 8.  | PRON       | 1942             | 8,993239 |
| 9.  | Verb       | 1749             | 8,099472 |
| 10. | Verb~CONV  | 1367             | 6,330462 |
| 11. | Verb~PTCPL | 293              | 1,356858 |

|               |       |              |             |
|---------------|-------|--------------|-------------|
| 12.           | CONV  | 1663         | 7,701213    |
| 13.           | PTCPL | 978          | 4,529036    |
| 14.           | CONJ  | 434          | 2,009818    |
| 15.           | POST  | 82           | 0,379735    |
| 16.           | PART  | 664          | 3,074928    |
| <b>Итого:</b> |       | <b>21594</b> | <b>100%</b> |

Из таблицы видно, что лидирующее положение по частоте употребления занимают имена существительные (собственные и нарицательные), которые встретились в тексте памятника 9 095 раз. Глагол вместе с омонимичными глагольными формами (Verb + Verb~Conv + Verb~PTCPL) занимает второе место по частоте употребления. Случаев морфологической омонимии, как показывает материал, достаточно много (2 074), что вызывает трудности в его анализе.

Итак, в ходе анализа «раннего» текста выявлен ряд проблем, касающихся транслитерации текста «тодо бичиг», орфографии текста, омонимии словоформ, разметки текста, использования диакритических знаков и т. д. Эти проблемы будут учтены впоследствии при обработке массива текстов на «тодо бичиг» для включения их в Национальный корпус калмыцкого языка. Необходимо реализовать следующие шаги для обработки текстов на «тодо бичиг»:

- видоизменить правила транслитерации текстов на «тодо бичиг»;
- создать грамматический словарь старокалмыцкого текста;
- создать словарь омонимичных форм на старокалмыцком языке;
- создать словарь вариантов написания одного и того же слова;
- создать словарь грамем старокалмыцкого языка.

Отдельной задачей стоит создание программы, распознающей старокалмыцкую письменность, а также конвертера тодо бичиг на латиницу/кириллицу, реализация которых действительно необходима для увеличения «ранних» текстов в корпусе калмыцкого языка, поскольку их изучение носит ретроспективный характер и охватывает самый широкий круг вопросов — от текстологии и диалектологии до сравнительно-исторического изучения словоформ, словосочетаний и т. д. Это может привести в свою очередь к реконструкции ойратских и общемонгольских древностей на вербальном уровне. Нередко в ранних текстах фик-

сируются лексемы и целые последовательности лексем, которые не встречаются в современных данных языка.

#### Литература

- Бадагаров Ж. Б.* О репрезентативности текстов и элементах программного инструментария для корпуса бурятского языка // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. Е1' Manuscript-08. Материалы Международной научной конференции (Казань, 26–30 августа 2008 г.). Казань, 2008. С. 28–31.
- Бадмаева Л. Д., Бадагаров Ж. Б., Цыдыпов Б. З.* Общие проблемы формирования корпуса бурятского языка // Труды Международной конференции «Корпусная лингвистика – 2008». СПб.: Изд-во Фил. фак-та СПбГУ, 2008. С. 24–30.
- Долинский В. А.* Квантитативная лингвистика в исследовании текста // Алфавит: Строеие повествовательного текста. Синтагматика. Парадигматика. Смоленск: СГПУ, 2004. С. 283–324.
- Зализняк А. А.* Словоформа // Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой; Ин-т языкознания АН СССР. М.: Сов. энцикл., 1990. С. 470.
- Крылов С. А.* Структурно-вероятностная модель современного монгольского языка (на базе Генерального корпуса монгольского языка) // Урало-алтайские исследования. 2012. № 1(6). С. 78–105.
- Куканова В. В.* Архитектура метаописания в Национальном корпусе калмыцкого языка // Вестник Калмыцкого института гуманитарных исследований РАН. 2011а. № 1. С. 139–145.
- Куканова В. В.* Общая структура и перспективы использования Национального корпуса калмыцкого языка в свете проблемы репрезентативности // Материалы XL Международной филологической конференции (10–15 марта 2011 г., Санкт-Петербург). Вып. 14: Полевая лингвистика и интегральное моделирование речи / отв. ред. Н. В. Богданова. СПб.: Изд-во Фил. фак-та СПбГУ, 2011б. С. 125–137.
- Леннгрен Л. (ред.).* Частотный словарь современного русского языка. Uppsala, 1993. 188 с.
- Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка).



М.: Азбуковник, 2009. [электронный ресурс] // <http://dict.ruslang.ru/freq.php> (дата обращения: 15.04.2012).

Ринчинов О. С. Корпус бурятского языка и прикладные задачи компьютерной лингвистики // Состояние и перспективы развития бурятского языка. Мат-лы форума бурятского языка. Улан-Удэ, 2009. С. 88–89.

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

Шайкевич А. Я. Количественные методы в языкознании // Лингвистический энциклопедический словарь / под ред. В. Н. Ярцевой; Ин-т языкознания АН СССР. М.: Сов. энцикл., 1990. С. 231.

УДК 81'33

ББК 81.23

## СЛОВОИЗМЕНТЕЛЬНЫЕ ТИПЫ В КАЛМЫЦКОМ ЯЗЫКЕ В СВЕТЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ (на примере имени существительного)

В. В. Куканова

Как известно, создание корпуса того или иного языка и его развитие является одной из актуальнейших задач современной лингвистики, в частности и калмыцкого языкознания, поскольку решается судьба языка как средства познания и коммуникации. Для того чтобы корпус «заработал», требуется пройти несколько шагов, связанных не только со сбором репрезентативного текстового материала, его филологической выверки. Одной из первых лингвистических задач является создание морфологического анализатора — программы, приводящей словоформу к начальной форме с приписыванием ей той или иной грамматической информации (грамматем), которую несет данная единица в тексте.

Для реализации этого программного продукта необходима разработанная система словоизменительных типов. Если речь идет о флективных языках, то выделение словоизменительных типов не вызывает никаких недоумений со стороны лингвистов. Система склонения и спряжения здесь более разнообразна и отличается нестандартностью выражения тех или иных граммем. Так, например, для русского языка насчитывается около 10 глагольных словоизменительных классов [Русская грамматика 1980: 647–661] и несколько видов аффиксов для выражения одной и той же граммемы. Если же говорить об агглютинативных языках, каким является калмыцкий язык, то имеются «сложности» выделения словоизменительных типов, так как они здесь более регулярны, единообразны и стандартны [Булыгина, Крылов 1990].

Во-первых, главной особенностью подобных языков является теоретическая возможность присоединения в строгом порядке неограниченного количества слово-

изменительных аффиксов и частиц к основе слова. Во-вторых, определенный тип парадигмы выделяют на основе общей системы выражения грамматических категорий определенными словоизменительными аффиксами и наличия сходных морфонологических процессов, а также частеречной принадлежности, в пределах которой действует тот или иной образец склонения (к примеру, в русском языке существуют субстантивное, адъективное и местоименное склонения)<sup>1</sup>.

В калмыцком языке имеет место четкое противопоставление именного и глагольного словоизменения: для каждого из них существуют неомонимичные аффиксы (сразу же оговоримся: речь идет только о формантах, выражающих определенные грамматические категории имени и глагола). Для всех именных частей речи — существительных, местоимений (за исключением некоторых форм с супплетивными основами), числительных — используются те же самые словоизменительные аффиксы для выражения категорий числа<sup>2</sup>, падежа и др. По сути,

<sup>1</sup> Е. С. Кубрякова и П. А. Соболева выделяют пять следующих обязательных признаков морфологической парадигмы: 1 — «наличие константной лексемы, выражающей в парадигме идею тождества слова самому себе во всех его видоизменениях, и переменных формантов»; 2 — «наличие константного набора грамматических значений, облигаторного для каждой из фиксируемых форм парадигмы и повторяемого от формы к форме с обязательным изменением одного из них»; 3 — «фиксированное количество переменных грамматических значений»; 4 — «для выражения каждого из грамматических значений (или набора) в парадигме <...> специальный формант»; 5 — «отношения производности, фиксируемые парадигмой» (например, косвенные падежи относительно исходной формы) [Кубрякова, Соболева 1979: 9].

<sup>2</sup> Категория числа в халха-монгольском языке, по мнению С. А. Крылова, схожа с категорией уменьши-