

КАЛМЫЦКИЕ СКАЗКИ, ЗАПИСАННЫЕ Г. Й. РАМСТЕДТОМ: ОПЫТ КВАНТИТАТИВНОГО АНАЛИЗА*

**Kalmyk Tales, Recorded by G. J. Ramstedt:
Experience of Quantitative Analysis**

В. В. Куканова (V. Kukanova)¹, Б. Б. Горяева (B. Goryaeva)²

¹ кандидат филологических наук, заведующий отделом теоретической и экспериментальной лингвистики Калмыцкого института гуманитарных исследований РАН (Ph. D. of Philology, Head of Laboratory of Theoretical and Experimental Linguistics at the Kalmyk Institute for Humanities of the RAS). E-mail: vika.kukanova@gmail.com.

² кандидат филологических наук, старший научный сотрудник отдела фольклора и джангароведения Калмыцкого института гуманитарных исследований РАН (Ph. D. of Philology, Senior Scientist of Folklore and Dzhangar Studies Department at the Kalmyk Institute for Humanities of the RAS). E-mail: baira79@yandex.ru.

Статья посвящена проблеме создания частотных словарей на материале калмыцкого языка. С помощью компьютерных технологий авторами создан частотный список, в котором эксплицированы тексты калмыцких сказок, записанных финским ученым Г. Й. Рамстедтом во время его путешествия в степи Нижнего Поволжья. С помощью методов квантитативного анализа выявлены частотные и нечастотные элементы калмыцкого языка в тексте, которые отбирались по нескольким статистическим показателям: абсолютной частоте словоупотребления, общей частоте, по коэффициентам D и R. Приведенные количественные показатели можно использовать для создания структурно-вероятностной модели языка. Актуальность создания подобной модели несомненна. Частотные списки позволяют определить границы лексической системы языка сказок, которая имеет свое ядро и периферийные поля, а также использовать в прикладных целях: с наиболее частотных единиц, как правило, начинается обучение языку, объясняется каково их значение и как использовать их в речи, что в свою очередь создаст теоретическую базу для распределения лексем по уровням изучения языка.

Ключевые слова: фольклор, сказка, калмыцкий язык, частотность, частотный словарь, квантитативная лингвистика, квантитативный анализ, Г. Й. Рамстедт.

The article is devoted to the creation of frequency dictionaries on the material of the Kalmyk language. With the help of computer technology, the authors created a frequency list in which they explicated the texts of Kalmyk tales, recorded by Finnish scholar G. J. Ramstedt during his travels in the steppes of the Lower Volga Region. Using the methods of quantitative analysis, frequent and non-frequent elements of the Kalmyk language, collected by several statistical indexes (absolute frequency of word usage, common frequency, the coefficients D and R) were identified in the text. Quantitative indexes can be used to create a structural-probabilistic model of language. The relevance of such model is undeniable. Frequency lists allow you to define the boundaries of the lexical system of fairy tales, which has its own core and peripheral fields, and to use for applied purposes: since teaching begins with the most frequent units and explaining their meaning and how to use them in speech, it will create a theoretical base for the distribution of tokens according to the levels of language learning.

Keywords: folklore, tale, the Kalmyk language, frequency, frequency dictionary, quantitative linguistics, quantitative analysis, G.J. Ramstedt.

В последние годы с развитием информационных технологий становится легче и быстрее создавать частотные списки, в которых эксплицирована структурно-вероятностная модель того или иного языка, определенного текста или подборки

текстов определенного жанра или автора. Методы квантитативной лингвистики приобретают все больший интерес среди исследователей, поскольку результаты количественной обработки текстов можно применить в решении не только приклад-

* Статья подготовлена при финансовой поддержке РГНФ, проект № 15-14-08002/а(р) «Фольклорный текст сквозь призму компьютерных технологий (на материале записей калмыцких сказок Г. Рамстедта)».

ных задач, но и фундаментальных теоретических проблем. Частотный словарь «... 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].

Создание частотных словарей на материале русского языка имеет уже продолжительную историю, начиная с 1950-х гг. [см.: Лёнгрен 1993; Степанова 1976; Частотный словарь ... 1977]. Венцом развития отечественной квантитативной лингвистики, конечно, стал Частотный словарь, основанный на материале Национального корпуса русского языка [Ляшевская, Шаров 2009], который насчитывал на момент работы над словарем 100 млн словоупотреблений¹. Именно этот словарь послужил нам примером и ориентиром в работе над созданием нашего Списка.

В данной работе мы рассмотрим результаты квантитативного анализа сказочных текстов, записанных Г.И. Рамstedтом в начале XX в. Во время путешествия в калмыцкие степи исследователь собрал обширный фольклорный материал: 22 сказки, около 100 загадок, 200 пословиц, 40 народных песен. Кроме того, он «записал 20 фонограмм калмыцких мелодий, а также сделал множество фотографических снимков из жизни калмыков» [Отчет ... 1904: 13].

Калмыцкие сказки были изданы в двух частях. Первая часть, состоящая из 154 страниц и содержащая 18 сказок, была издана в 1909 г. [Kalmückische ... 1909]. Вторую часть (с. 155–237) ученый планировал издать на следующий год, но она увидела свет только в 1919 г. и содержала сказки под номерами 19–22 [Kalmückische ... 1919]. Обе части издания калмыцких сказок вышли без предисловия. Г. И. Рамstedт собирался издать серию книг, посвященных калмыцкому фольклору: сказки, пословицы, поговор-

ки, песни и эпические сказания. Он намеревался собрать дополнительный материал по калмыцкому фольклору в последующих своих экспедициях к волжским калмыкам, однако этого не удалось сделать: фольклорные записи увидели свет в разрозненных изданиях (журнальном и книжном вариантах).

Сказки, зафиксированные Г.И. Рамstedтом, были набраны одним из авторов статьи (А. О. Бутаевой) в фонетической транскрипции, затем автоматически сегментированы на предложения, которые были соотнесены с переводом на немецкий язык (3 192 параллельных блока). Записи были набраны в соответствии с концепцией, принятой разработчиками фольклорного подкорпуса² [Куканова, Манджиева, Горяева 2013]. Тексты были переложены Б. Б. Горяевой на современную калмыцкую орфографию с сохранением диалектной лексики.

Главная идея создания частотных списков заключается в том, что наиболее частотная единица является наиболее важной в системе, и в то же время нечастотные элементы свидетельствуют о неупотребительности лексической единицы в речи и занимают уникальное место в лексической системе. Они могут содержать следы исторического развития, принадлежать той или иной терминологической системе.

«Верхушки» созданных частотных списков словоформ и лемм приведены ниже в таблицах. Поскольку принят за основу Частотный словарь, основанный на материале Национального корпуса русского языка [Ляшевская, Шаров 2009], мы взяли следующие статистические показатели.

1. Абсолютная частота словоупотребления (2).

2. Ipm (instances per million words), общая частота употребления на миллион слов условного корпуса (4).

3. Ранг (R), который присваивается в зависимости от частоты словоупотребления в текстах. Так, самое частотное слово получает ранг 1, следующее — ранг 2 и т. д. (5).

4. Коэффициент D (введен А. Жюйяном) для определения показателя равномерности распределения того или иного слова в различных текстах корпуса в пределах интервала 0–100 (3).

5. Показатель R (%) для определения количества текстов, в которых встретилось слово (6).

² В рамках Национального корпуса калмыцкого языка (kalmcorporu.ru).

¹ Ср. с частотным словарем под ред. Л. Н. Засориной, который основан на текстах общим объемом 1 млн словоупотреблений [Частотный словарь ... 1977].

1. «Верхушка» частотного списка словоформ, отсортированного по абсолютной частоте и рангу

Текст	Частота	D	IPM	R	R(%)
1	2	3	4	5	6
гигж	691	82	16655	1	90
болна	526	66	12678	2	59
нег	503	79	12123	3	100
гигһәд	394	73	9496	4	100
гинә	347	71	8363	5	68
авад	331	83	7978	6	100
келв	331	74	7978	6	95
уга	306	75	7375	8	90
көвүн	300	69	7230	9	77
өвгән	298	68	7182	10	72
хойр	267	78	6435	11	95
һарад	254	82	6122	12	86
ирәд	211	84	5085	13	100
күн	200	79	4820	14	90
күрәд	183	79	4410	15	72
эн	179	80	4314	16	81
ирнә	179	70	4314	16	72
эмгнә	169	79	4073	18	72
бәәнә	147	71	3543	19	77
тер	145	78	3494	20	86
күүкән	145	66	3494	20	63
би	142	73	3422	22	86
һурвн	141	79	3398	23	86
хар	140	69	3374	24	54
авч	132	82	3181	25	86
ик	129	79	3109	26	77
бәәдг	129	65	3109	26	54
чигн	125	66	3012	28	68
санж	124	86	2988	29	90
болад	121	76	2916	30	63
ирв	120	77	2892	31	77
хан	119	68	2868	32	59
йовна	115	74	2771	33	77
келжәнә	114	73	2747	34	77
йовад	110	76	2651	35	77
күрч	108	77	2603	36	90
хаана	107	69	2578	37	68
келәд	100	79	2410	38	77

Самой частотной словоформой в сказках является слово *гуж*, что не случайно. Если сравнить с данными по частотам в русском языке, то одним из частотных слов является союз *что*. Однако в калмыцком языке словоформа *гуж* является грамматическим омонимом и может выступать как соеди-

нительное деепричастие от глагола *гых* 'говорить' и как союз. Отметим, что в список частотных единиц попали словоформы, которые обозначают речевые действия: *гунэд*, *гинэ*, *келв*, *келжэнэ*, *келэд*. Среди частотных глаголов можно обнаружить глаголы движения: *һарад*, *ирэд/ирв*, *йовад*.

2. «Верхушка» частотного списка словоформ, отсортированного по коэффициенту *D*

Текст	Частота	D	IPM	ABS	R	R(%)
<i>санж</i>	124	86	2988	20916	29	90
<i>ирэд</i>	211	84	5085	35595	13	100
<i>авад</i>	331	83	7978	55846	6	100
<i>авч</i>	132	82	3181	22267	25	86
<i>гуж</i>	691	82	16655	116585	1	90
<i>һарад</i>	254	82	6122	42854	12	86
<i>одв</i>	86	81	2072	14504	47	77
<i>амр</i>	22	80	530	3710	216	68
<i>кезэнэ</i>	27	80	650	4550	168	77
<i>эн</i>	179	80	4314	30198	16	81
<i>деер</i>	64	80	1542	10794	61	68
<i>тиигж</i>	33	80	795	5565	134	72
<i>нег</i>	503	80	12123	84861	3	100
<i>ик</i>	129	79	3109	21763	26	77
<i>күрэд</i>	183	79	4410	30870	15	72
<i>эмгнэ</i>	169	79	4073	28511	18	72
<i>һурвн</i>	141	79	3398	23786	23	86
<i>келэд</i>	100	79	2410	16870	38	77
<i>хойр</i>	267	79	6435	45045	11	95
<i>мини</i>	77	79	1855	12985	54	63
<i>күн</i>	200	79	4820	33740	14	90
<i>залу</i>	68	79	1638	11466	57	72
<i>гертэн</i>	63	78	1518	10626	63	77
<i>тер</i>	145	78	3494	24458	20	86
<i>чи</i>	81	78	1952	13664	49	72
<i>һазр</i>	55	78	1325	9275	72	68
<i>бээж</i>	31	78	747	5229	146	68
<i>күрч</i>	108	77	2603	18221	36	90
<i>ирв</i>	120	77	2892	20244	31	77
<i>тегэд</i>	81	77	1952	13664	49	72
<i>жсирһэд</i>	12	77	289	2023	415	50
<i>бэрж</i>	33	76	795	5565	134	54
<i>йовад</i>	110	76	2651	18557	35	77
<i>болад</i>	121	76	2916	20412	30	63

хэрж	13	76	313	2191	379	50
уга	306	75	7375	51625	8	90
орад	89	75	2145	15015	45	81
хот	18	75	433	3031	268	50
юмн	60	75	1446	10122	68	68
алад	53	75	1277	8939	75	72
келв	331	75	7978	55846	6	95
келж	31	75	747	5229	146	54
болж	28	74	674	4718	164	54
кеһад	45	74	1084	7588	92	63
одна	49	74	1181	8267	77	77
һаза	18	74	433	3031	268	54
хэрэд	15	74	361	2527	318	50
гунһад	394	74	9496	66472	4	100
йовна	115	74	2771	19397	33	77
тана	22	74	530	3710	216	50
һарһад	26	74	626	4382	178	50
орж	34	74	819	5733	129	59
ю	68	74	1638	11466	57	72
би	142	73	3422	23954	22	86
келжәнә	114	73	2747	19229	34	77
нанд	48	73	1156	8092	81	68
юн	95	73	2289	16023	41	68
тиигхлә	76	73	1831	12817	55	72
һарна	47	73	1132	7924	86	59
гер	24	72	578	4046	195	54
улан	65	72	1566	10962	60	50
тал	32	72	771	5397	138	50
чамд	22	72	530	3710	216	54
ардаснь	25	72	602	4214	189	54
бэв	60	72	1446	10122	68	63
босад	35	72	843	5901	123	50
ав	29	72	698	4886	154	54
авна	29	72	698	4886	154	54
бичг	42	72	1012	7084	101	54
йовсн	44	72	1060	7420	94	63
бэсн	50	72	1205	8435	76	63
олж	23	72	554	3878	206	45
гихлә	32	71	771	5397	138	59
цокад	35	71	843	5901	123	50
менд	43	71	1036	7252	97	54
кевтән	26	71	626	4382	178	54

унад	13	71	313	2191	379	40
бээнэ	147	71	3543	24801	19	77
дахулад	19	71	457	3199	251	45
уунад	19	71	457	3199	251	45
йовэж	63	71	1518	10626	63	63
хөр	21	71	506	3542	229	40
гинэ	347	71	8363	58541	5	68
мана	43	71	1036	7252	97	68
авн	20	71	482	3374	239	45
ир	28	71	674	4718	164	54
ирж	36	71	867	6069	115	59

В таблице приведены данные по распределению словоформ относительно друг друга в линейной последовательности. Отсутствует единица, которая занимала бы

стопроцентную позицию по равномерному распределению. Более или менее равномерно распределены указанные в таблице словоформы.

3. «Верхушка» частотного списка словоформ, отсортированного по коэффициенту R

Текст	Частота	D	IPM	ABS	R	R(%)
нег	503	80	12123	84861	3	100
гилэд	394	74	9496	66472	4	100
авад	331	83	7978	55846	6	100
ирэд	211	84	5085	35595	13	100
келв	331	75	7978	55846	6	95
хойр	267	79	6435	45045	11	95
гилж	691	82	16655	116585	1	90
уга	306	75	7375	51625	8	90
күн	200	79	4820	33740	14	90
санж	124	86	2988	20916	29	90
күрч	108	77	2603	18221	36	90
һарад	254	82	6122	42854	12	86
тер	145	78	3494	24458	20	86
би	142	73	3422	23954	22	86
һурвн	141	79	3398	23786	23	86
авч	132	82	3181	22267	25	86
эн	179	80	4314	30198	16	81
орад	89	75	2145	15015	45	81
һарч	67	67	1614	11298	59	81
көвүн	300	69	7230	50610	9	77
бээнэ	147	71	3543	24801	19	77
ик	129	79	3109	21763	26	77
ирв	120	77	2892	20244	31	77
йовна	115	74	2771	19397	33	77
келжэнэ	114	73	2747	19229	34	77

йовад	110	76	2651	18557	35	77
келад	100	79	2410	16870	38	77
окна	89	71	2145	15015	45	77
одв	86	81	2072	14504	47	77
ода	82	70	1976	13832	48	77
гертан	63	78	1518	10626	63	77
одна	49	74	1181	8267	77	77
кезэнэ	27	80	650	4550	168	77
өвгэн	298	69	7182	50274	10	72
курэд	183	79	4410	30870	15	72
ирнэ	179	70	4314	30198	16	72
эмгнэ	169	79	4073	28511	18	72
чи	81	78	1952	13664	49	72
тегэд	81	77	1952	13664	49	72
тиигхлэ	76	73	1831	12817	55	72
залу	68	79	1638	11466	57	72
ю	68	74	1638	11466	57	72
алад	53	75	1277	8939	75	72
тиигжэ	33	80	795	5565	134	72

Данный показатель R (так называемая покрываемость текста лексической единицей) указывает на то, насколько часто та или иная единица используется в разных текстах. Возьмем, к примеру, словоформы

нег, гинэд, авад, ирэд, которые используются во всех 22 текстах сказок, что свидетельствует о том, что данные единицы носят универсальный характер.

4. «Верхушка» частотного списка лемм, отсортированного по абсолютной частоте словоупотребления

R	Текст	Частота
1	гнх	1672
2	болх	985
3	келх	847
4	ирх	757
5	авх	676
6	йовх	593
7	негн	540
8	нарх	473
9	бээх	395
10	көвүн	368
11	би	340
12	уга	336
13	хан	325
14	курх	317
15	хойр	317

16	өвгн	578
17	күн	313
18	одх	288
19	бээх	264
20	өгх	218
21	эмгн	213
22	орх	208
23	күүкн	204
24	тер	192
25	хурвн	187
26	юн	174
27	эн	172
28	окх	166
29	чи	161
30	кех	153
31	эмгн	149

33	<i>алх</i>	148
34	<i>хар</i>	140
35	<i>сурх</i>	135
36	<i>ик</i>	131
37	<i>идх</i>	130
38	<i>чигн</i>	125

39	<i>махн</i>	120
40	<i>куукн</i>	116
41	<i>һазр</i>	110
42	<i>ах</i>	102
43	<i>юмн</i>	102

Как и в русском языке, самыми частотными единицами в калмыцких сказках является речевой глагол *гилх* ‘говорить’ (1672), полужнаменательный глагол *болх* ‘становиться’ (985), речевой глагол *келх* ‘говорить, сказать’ (757), глаголы движения *ирх* ‘прийти’ (757) и *йовх* ‘идти’, *авх* ‘брат’ (676). Среди существительных можно выделить *овгн* ‘старик’ (578), *көвүн* ‘мальчик’ (368), *хан* ‘хан’ (325) и др.

Таким образом, приведенные количественные показатели можно использовать для создания структурно-вероятностной модели языка сказочных текстов, записанных

Г. Й. Рамстедтом. Актуальность создания подобной модели несомненна. Во-первых, частотные списки позволяют определить границы лексической системы языка сказок, которая имеет свое ядро и периферийные поля (т. е. частотные и нечастотные элементы). В аспекте практической значимости создание частотных списков можно использовать в прикладных целях: с наиболее частотных единиц, как правило, начинается обучение языку, объясняется, каково их значение и как использовать их в речи, что в свою очередь создаст теоретическую базу для распределения лексем по уровням изучения языка.

Литература

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

Куканова В. В., Манджиева Б. Б., Горяева Б. Б. Оцифровка фольклорных произведений: вызовы и решения // Известия Высших учебных заведений. Северо-Кавказский регион. 2013. № 6. С. 123–129.

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. 1112 с.

Лёнгрен Л. (ред.). Частотный словарь современного русского языка [Lönngren, Lennart. The Frequency Dictionary of Modern Russian. Acta

Univ. Ups., Studia Slavica Upsaliensia Uppsala 32]. Uppsala, 1993.188 с.

Отчет д-ра Г.Й. Рамстедта за 1903 год // Известия Русского комитета для изучения Средней и Восточной Азии. 1904. №2. С. 11–14

Степанова Е. М. Частотный словарь общенаучной лексики. М.: Просвещение, 1976. 87 с.

Частотный словарь русского языка / под ред. Л. Н. Засориной. М.: Русский язык, 1977. 936 с.

Kalmückische Sprachproben. Gesammelt und herausgegeben von G. J. Ramstedt. Erster Teil. Kalmückische Märchen. – Helsingfors: Societe Finno-Ougrienne, 1909. 154 s.

Kalmückische Sprachproben. Gesammelt und herausgegeben von G. J. Ramstedt. Zweiter Teil. Kalmuckische Märchen. Helsingfors: Societe Finno-Ougrienne, 1919. S. 155–237.

References

[Frequency Dictionary of the Russian Language]. L. N. Zazorina (ed.). Moscow: Russkiy Yazyk, 1977. 936 p. (In Russ.)

[Kalmyk Speech Patterns. Collected and Published by G. J. Ramstedt]. Part 1: Kalmyk Tales. – Helsinki: Finno-Ugric Society, 1909. 154 p. (In Kalm. and Germ.)

[Kalmyk Speech Patterns. Collected and Published by G. J. Ramstedt]. Part 2: Kalmyk Tales. Helsinki: Finno-Ugric Society, 1919. Pp. 155–237. (In Kalm. and Germ.)

1903 Yearly Report by Dr. G. J. Ramstedt. *Izvestiya Russkogo komiteta dlya izucheniya Sredney i Vostochnoy Azii*. 1904. No. 2. Pp. 11–14. (In Russ.)

Dolinsky V. A. Quantitative linguistics in textual studies. In: [Alphabet: Structure of a Narrative

Text. Syntagmatics. Paradigmatics]. Smolensk: Smolensk State Pedagogical University, 2004. Pp. 283–324. (In Russ.)

Kukanova V. V., Mandzhieva B. B., Goryaeva B. B. Digitizing folklore texts: challenges and solutions. *Bulletin of Higher Institutes. North Caucasus Region*. 2013. No. 6. Pp. 123–129. (In Russ.)

Lönngren L. The Frequency Dictionary of Modern Russian. Uppsala, 1993.188 p. (In Russ. and Swed.)

Lyashevskaya O. N., Sharov S. A. [Frequency Dictionary of Modern Russian: a Case Study of Russian National Corpus]. Moscow: Azbukovnik, 2009. 1112 p. (In Russ.)

Stepanova E. M. [Frequency Dictionary of Common Scientific Terms]. Moscow: Prosveshchenie, 1976. 87 p. (In Russ.)