Известия Гомельского государственного университета
имени Ф. Скорины, №2(41), 2007

Языкознание

УДК 81.32

Частотные характеристики лексического состава паремий

Н. И. Ефремова

Многие выводы о лексическом наполнении пословиц и поговорок можно получить с помощью статистических методов. «Подобные содержательные суждения, – отмечает Р.М.Фрумкина, – основанные на точном количественном критерии, имели бы большое преимущество перед теми утверждениями о словарном составе текстов, которые основаны на субъективном представлении о «богатстве» словаря, о преобладании каких-либо классов слов или форм» [8, с. 45]. Для проведения такого анализа нами были составлены частотные словари лексики исследуемых корпусов русских и немецких пословиц и поговорок, а также построены таблицы распределения частот (статистические структуры). Материалом для исследования послужили 1000 русских и 1000 немецких паремий [2, 3, 4].

1 Частотные словари лексики немецких и русских паремий

Частотный словарь представляет собой список лексем, расположенных в алфавитном или ином определенном порядке, в котором «каждая входная единица сопровождается указанием на частоту ее употребления в тексте, использованном для составления этого словаря» [1, с. 7]. В настоящем исследовании частотные словари пословичных макротекстов составлялись путем подсчета всех словоупотреблений знаменательных и полузнаменательных слов. За словарную единицу принята лексема. Все лексемы в обоих словарях представлены в исходных формах и расположены по убывающим частотам. Слова, имеющие одинаковую частоту, расположены в алфавитном порядке и пронумерованы в соответствии с ним.

1.1 Принципы идентификации лексем

Идентификация лексем производилась с учетом правил, сформулированных отдельно для каждой части речи конкретного языка. Так, при анализе существительных составные единицы разбивались на простые: *хлеб-соль*, *Sankt Peter* ‘Святой Петр’;

– уменьшительные имена объединялись с соответствующими исходными лексемами: *хлеб – хлебушко*, *Hase* ‘заяц’ – *Häschen* ‘зайчик’;

– звательные формы имен существительных возводились к именительному падежу: *боже – бог*;

– фонетические варианты слов сводились к инварианту лексем: *глас – голос*, *lange – lang* ‘долго’;

– супплетивные формы мн. ч. существительных объединялись с соответствующими формами ед. ч.: *человек – люди*, *Mann – Leute* ‘человек – люди’;

– в частотный словарь включены все имена собственные, встретившиеся в анализируемых пословичных текстах.

Глагольные формы (личные, безличные, причастные, деепричастные, а также причастие I и причастие II немецких глаголов) сводились к инфинитиву: *учась – учиться*, *rede – reden* ‘говори – говорить’;

– глаголы разного залога и вида считались разными лексемами: *взять*, *взяться*, *лить*, *пролить*, *gehen* ‘идти’, *vergehen* ‘проходить’;

– редуцированные формы немецких глаголов и усеченные формы русских объединялись с их полными вариантами: *beginn' – beginnen* ‘начинать’, *besinn' – besinnen* ‘вспоминать’, *щелк – щелкнуть*.

Прилагательные и наречия, имеющие степени сравнения, представлены в словаре в положительной степени; супплетивные формы считались разновидностью исходных лексем: *хорошо – лучше*, *viel* ‘много’ – *mehr* ‘больше’;

– немецкие омонимичные прилагательные и наречия представлены в частотном словаре раздельно: *früh* ‘ранний’, *früh* ‘рано’.

– Местоимения, имеющие формы рода, числа и падежа, зафиксированы в словаре в им. п. ед. ч. мужского рода: *мой*, *der* ‘который’;

– начальной формой для личных местоимений является им. п. единственного числа соответствующего лица: *мне – я*, *dir – du* ‘тебе – ты’;

– местоимения *каков*, *таков* зафиксированы в словаре как фонетические варианты лексем *какой*, *такой*.

Количественные, порядковые и собирательные числительные рассматривались как самостоятельные разряды слов; они даны в начальных формах: *два*, *второй*, *двое*, *drei*, *dritte*, *dreie* ‘три, трое, третий’.

В результате подсчета всех словоупотреблений каждой лексемы получены абсолютные частоты лексем и составлены частотные словари лексики.

1.2 Оценка релевантности (адекватности) частотных словарей

Чтобы производить дальнейшие действия с единицами частотных словарей, необходимо доказать, что они могут быть использованы в качестве достоверных и репрезентативных источников данных, релевантных для совокупного корпуса русских паремий и совокупного корпуса немецких паремий.

Эмпирической проверкой надежности частотного словаря служит оценка его релевантности или адекватности. Релевантность словаря проверяется при сравнении его данных с данными других словарей, меньших по объему. В качестве контрольной выборки частотного словаря паремийной лексики взят сборник «Русские пословицы и поговорки и их немецкие аналоги» [5]. Словарь включает 432 русские и 568 немецких пословиц и поговорок и содержит список лексем с указанием их частоты. В качестве критерия для оценки адекватности проводится сопоставление двадцати самых частых слов анализируемых источников.

Как показал анализ, 15 из двадцати высокочастотных немецких слов контрольной выборки имеют соответствия в основном материале. К их числу принадлежат лексемы *sein*, ‘быть’, *man*, *haben*, *wer*, *der*, *es*, *gut*, ‘хорошо’, *machen*, *alle/s*, *er*, *können*, *müssen*, *sein*’его’, *werden*, *ander*, причем слова *sein*, ‘быть’, *man*, *haben* сохранили свои ранговые положения (соответственно I, II, III ранги). Русские высокочастотные слова, представленные в сборнике «Русские пословицы и поговорки и их немецкие аналоги», также повторялись в высокочастотной зоне исследуемого материала. Группа включает лексемы *все*, *бог*, *один*, *хорошо*, *голова*, *всякий*, *дело*, *жить*, *вода*, *ум*, *два*, *дурак*, *волк*. Одинаковый частотный ранг в двух списках сохранила лексема *бог*.

Таким образом, 75% из двадцати самых частых слов немецкого материала и 65% из двадцати высокочастотных лексем русского материала имеют соответствия в контрольной выборке, что позволяет говорить об адекватности (релевантности) составленных частотных словарей по отношению к лексике генеральной совокупности паремий.

1.3 Степень лексического разнообразия словарных составов пословичных текстов

Важным статистическим параметром частотных словарей некоторого текста является соотношение количества словоупотреблений в тексте и количества разных лексем. Данное соотношение трактуется в статистике текста как коэффициент (степень) разнообразия словарного состава.

Исследуемая выборка русских паремий состоит из 4022 текстовых единиц и содержит 1677 разных лексем, следовательно, коэффициент лексического разнообразия для русского материала составляет 0,42. Немецкие паремии имеют общую длину текстов в 4828 словоупотреблений при 1471 разной лексеме; коэффициент лексического разнообразия для них составил 0,31. Показателем разнообразия словарного состава текстов является и количество слов с частотой 1. Удельный вес лексики с абсолютной частотой 1 в словаре русских паремий составляет 26 %, для немецких паремий соответствующий показатель равен 18 %. Судя по коэффициенту разнообразия и удельному весу редких слов, немецкие паремии проявляют меньшее разнообразие лексического наполнения, что подтверждается и различиями в статистических структурах словарей лексики двух пословичных макротекстов.

Русские пословицы и поговорки, являясь в своем большинстве образными, передают житейские истины, социальные нормы иносказательно, используя в качестве образной основы богатый материал из природы, быта, повседневной жизни людей. Буквальный план паремий создается в основном в результате употребления низкочастотных лексем, например: *помело* (2)[1], *обух* (2), *плюнуть* (1). Для многих немецких паремий характерно отсутствие образности. В создании обобщающей семантики пословиц и поговорок участвуют высокочастотные абстрактные существительные: *Glück* ‘счастье’(25), *Liebe* ‘любовь’(12), *Not* ‘нужда’(12), *Armut* ‘бедность’(11), *Recht* ‘право’(10), что, естественно, снижает степень разнообразия словаря.

Многие немецкие пословицы, содержащие предписания, наставления, поучения, выражают их, в отличие от русских паремий, не образно-иносказательно, а эксплицитно. Для этого используются модальные глаголы *müssen* (36), *sollen* (27), *lassen* (10), высокочастотность которых в целом снижает степень разнообразия лексического состава. Степень лексического разнообразия немецких паремий снижается также из-за присутствия высокочастотных личных местоимений: *er* (45), *du* (32), *ich* (29), *sie* (11). Сходное содержание русских пословиц и поговорок может передаваться без участия прономинативов: *Er hat läuten gehört, weiß aber nicht, wo die Glocken hängen* ‘Он слышал, что звонят, но не знает, где висят колокола’ – *Слышал звон, да не знает, где он.*

2 Статистическая структура словарей пословичной лексики

Чтобы получить общее представление о распределении частот в частотных словарях пословичной лексики, нами определены статистические структуры словарей, которые представляют собой таблицы с численными характеристиками входящих в них единиц (См. приложения 1 и 2).

2.1 Методы табулирования данных

Основными количественными показателями статистической структуры словаря являются абсолютная частота лексемы (F) и количество слов с определенной частотой (n). Произведение абсолютной частоты и количества слов с этой частотой (Fn) дает число всех словоупотреблений с заданными абсолютными показателями. Лингвостатистическая таблица может содержать и дополнительные (факультативные) сведения, как, например, накопленная абсолютная частота (F*), то есть количество всех словоупотреблений, зарегистрированных до определенной абсолютной частоты. Последняя цифра в колонке накопления абсолютных частот указывает на объем всего словаря.

Наряду с абсолютной частотой важным показателем статистической структуры словаря является относительная частота – «отношение абсолютной частоты к числу произведенных опытов или к числу единиц в обследованном массиве» [6, с.10]. Относительные частоты, помещенные в таблицах, округлены до пятого знака после запятой. Суммируя относительные частоты лексем, мы получаем накопленную относительную частоту (f*), которая позволяет судить о доле текста, занимаемой определенной группой слов.

---

[1] В скобках указаны абсолютные частоты лексем

Для фиксации количества разных частот в словаре вводится ранговая нумерация (r). Слова, имеющие одинаковую частоту, расположены в пределах одного ранга. Подсчет числа разных лексем осуществляется с помощью сплошной нумерации (N). Последний номер в колонке сплошной нумерации указывает на количество разных слов, помещенных в словаре. Все вычисления были произведены с помощью электронных таблиц EXEL.

2.2 Статистические характеристики разных зон частотных словарей пословичной лексики

В целях сопоставительного анализа частотных характеристик пословичных текстов статистическая структура каждого словаря условно разделена на три зоны: верхнюю, среднюю и нижнюю. Верхняя зона словаря немецких паремий охватывает частотные ранги 1 – 14, средняя зона – ранги 15 – 28, нижняя зона – ранги 29 – 42. Статистическая структура словаря русских паремий состоит из зон, соответствующих частотным рангам 1 – 11, 12 – 22, 23 – 32.

Как видно из приложения 1, верхняя зона частотного словаря русских паремий включает 14 разных слов, которые соответствуют значительному числу словоупотреблений – 482. Однако самые частые слова покрывают небольшую часть паремиологических текстов (11,98%). Абсолютные частоты слов варьируют в пределах от 73 до 22, однако количество слов с самыми высокими абсолютными показателями минимально: от одной до двух лексем. Согласно исследованиям Р.М.Фрумкиной, «100 наиболее частых слов покрывают в любом тексте очень большое количество словоупотреблений, порядка 70 - 75 %» [7, с. 74]. В нашем исследовании наблюдается обратная тенденция: самым частым словам отводится небольшая роль в формировании словаря текста, что свидетельствует о высокой степени разнообразия паремий.

Статистическая структура верхней зоны словаря немецких паремий существенно отличается от структуры верхней зоны словаря русских пословиц и поговорок. Большее количество частотных рангов в немецком материале предполагает большее количество разных слов, а именно – 16 лексем. Максимальная абсолютная частота в данном списке равна 246, минимальная – 36. Значительные расхождения в абсолютных частотах верхних зон немецких и русских словарей создаются за счет высокой фразообразовательной активности немецких лексем *sein*, *haben*, *man*, которые в большинстве контекстов выполняют вспомогательную функцию в составе предложения. Высокочастотные слова в немецком списке составляют 1088 словоупотреблений и покрывают 22,53% лексического наполнения паремиологических текстов, что в два раза превышает соответствующий показатель русских паремий. Большое количество высокочастотной лексики значительно снижает степень разнообразия словарного состава немецких пословиц и поговорок.

Средние зоны русских и немецких словарей характеризуются наименьшими различиями в своей структуре. Низкочастотная лексика расположена в нижних зонах частотных словарей. Она представляет особый интерес для исследования, так как за счет редких слов создается богатство словаря. Высокочастотная лексика и лексика средней частотности определяют основные тематические направления текстов, однако они выполняют незначительную роль в формировании лексического разнообразия пословиц и поговорок. Редким словам, напротив, отводится большая роль не только в создании лексического и тематического разнообразия паремий, но и в создании стилистического многообразия лексики пословичных текстов. Среди низкочастотной лексики встречаются как общеупотребительные слова, так и слова, специфические для данного жанра: архаизмы, лакуны, а также стилистически окрашенная лексика.

Нижняя зона частотного словаря русских паремий включает абсолютные частоты от 10 до 1 и объединяет 1627 разных лексем, что соответствует 3051 словоупотреблению или 75,86%. Последний ранг в словаре занимают слова с абсолютной частотой 1, которые «свидетельствуют о богатстве словаря автора, с одной стороны, и указывают на устаревшую или неустоявшуюся лексику, с другой» [9, с. 927].

Статистическая структура словаря нижней зоны немецких паремий не отличается существенно от структуры нижней зоны словаря русских паремий по абсолютным частотам.

Для немецких пословиц и поговорок эти показатели представляют собой шкалу от 14 до 1. Однако число лексем с зафиксированными частотами ниже, чем в русском материале, вследствие чего уменьшилось и общее количество словоупотреблений. Редкие слова покрывают 65,39% всех текстовых единиц. Нижней зоне частотного словаря немецких паремий свойственна меньшая степень лексического разнообразия, чем соответствующей русской, что отражается и на разнообразии всего словника.

Русские паремиологические тексты являются более разнообразными, чем другие тексты на русском языке, так как имеют статистическую структуру, во многом отличную от структуры других текстов (См [9]). Верхняя зона частотного словаря пословиц и поговорок занимает сравнительно небольшую часть всех текстовых словоупотреблений, оставляя пространство для более редких слов и повышая тем самым лексическое разнообразие и богатство словарного состава паремий.

Приложение 1.

Статистическая структура словаря лексики русских пословиц и поговорок, составленного для исследованного корпуса в 1000 паремий

R	N	F	n	Fn	F*	f	f*
1	1	73	1	73	73	0,01815	0,01815
2	2	55	1	55	128	0,01367	0,03182
3	3	42	1	42	170	0,01044	0,04227
4	4	38	1	38	208	0,00945	0,05172
5	5	37	1	37	245	0,00920	0,06091
6	6	35	1	35	280	0,00870	0,06962
7	7-8	29	2	58	338	0,01442	0,08404
8	9	26	1	26	364	0,00646	0,09050
9	10-11	25	2	50	414	0,01243	0,10293
10	12	24	1	24	438	0,00597	0,10890
11	13-14	22	2	44	482	0,01094	0,11984
12	15	21	1	21	503	0,00522	0,12506
13	16	20	1	20	523	0,00497	0,13003
14	17	19	1	19	542	0,00472	0,13476
15	18-19	18	2	36	578	0,00895	0,14371
16	20	17	1	17	595	0,00423	0,14794
17	21	16	1	16	611	0,00398	0,15191
18	22	15	1	15	626	0,00373	0,15564
19	23-29	14	7	98	724	0,02437	0,18001
20	30-33	13	4	52	776	0,01293	0,19294
21	34-41	12	8	96	872	0,02387	0,21681
22	42-50	11	9	99	971	0,02461	0,24142
23	51-57	10	7	70	1041	0,01740	0,25883
24	58-72	9	15	135	1176	0,03357	0,29239
25	73-93	8	21	168	1344	0,04177	0,33416
26	94-109	7	16	112	1456	0,02785	0,36201
27	110-139	6	30	180	1636	0,04475	0,40676
28	140-176	5	37	185	1821	0,04600	0,45276
29	177-239	4	63	252	2073	0,06266	0,51542
30	240-356	3	117	351	2424	0,08727	0,60269
31	357-633	2	277	554	2978	0,13774	0,74043
32	634-1677	1	1044	1044	4022	0,25957	1,00000

Приложение 2.

Статистическая структура словаря лексики немецких пословиц и поговорок, составленного для исследованного корпуса в 1000 паремий

R	N	F	n	Fn	F*	f	f*
1	1	246	1	246	246	0,05095	0,05095
2	2	106	1	106	352	0,02196	0,07291
3	3	90	1	90	442	0,01864	0,09155
4	4	69	1	69	511	0,01429	0,10584
5	5	62	1	62	573	0,01284	0,11868
6	6	61	1	61	634	0,01263	0,13131
7	7	58	1	58	692	0,01201	0,14333
8	8	56	1	56	748	0,01160	0,15493
9	9	55	1	55	803	0,01139	0,16632
10	10	46	1	46	849	0,00953	0,17585
11	11	45	1	45	894	0,00932	0,18517
12	12	42	1	42	936	0,00870	0,19387
13	13-14	40	2	80	1016	0,01657	0,21044
14	15-16	36	2	72	1088	0,01491	0,22535
15	17	35	1	35	1123	0,00725	0,23260
16	18	33	1	33	1156	0,00684	0,23943
17	19-20	32	2	64	1220	0,01326	0,25269
18	21-22	30	2	60	1280	0,01243	0,26512
19	23	29	1	29	1309	0,00601	0,27112
20	24	27	1	27	1336	0,00559	0,27672
21	25-28	25	4	100	1436	0,02071	0,29743
22	29	24	1	24	1460	0,00497	0,30240
23	30-31	21	2	42	1502	0,00870	0,31110
24	32	19	1	19	1521	0,00394	0,31503
25	33-34	18	2	36	1557	0,00746	0,32249
26	35-38	17	4	68	1625	0,01408	0,33658
27	39	16	1	16	1641	0,00331	0,33989
28	40-41	15	2	30	1671	0,00621	0,34610
29	42-46	14	5	70	1741	0,01450	0,36060
30	47-52	13	6	78	1819	0,01616	0,37676
31	53-59	12	7	84	1903	0,01740	0,39416
32	60-70	11	11	121	2024	0,02506	0,41922
33	71-87	10	17	170	2194	0,03521	0,45443
34	88-97	9	10	90	2284	0,01864	0,47307
35	98-112	8	15	120	2404	0,02486	0,49793
36	113-132	7	20	140	2544	0,02900	0,52692
37	133-164	6	32	192	2736	0,03977	0,56669
38	165-203	5	39	195	2931	0,04039	0,60708
39	204-271	4	68	272	3203	0,05634	0,66342
40	272-365	3	94	282	3485	0,05841	0,72183
41	366-602	2	237	474	3959	0,09818	0,82001
42	603-1471	1	869	869	4828	0,17999	1,00000

Abstract. The paper touches on frequency characteristics of the lexical units of the paremias of two unkindred languages (Russian und German). Applying statistical methods to the paremialogical material the author defines the degree of lexical variety of the russian and german proverbs and sayings.

Литература

1. Алексеев П.М. Статистическая лексикография (типология, составление и применение частотных словарей): Учебное пособие. – Л.: ЛГПИ, 1975. – 120 с.

2. Бинович Л.Е., Гришин Н.Н. Немецко-русский фразеологический словарь. Изд. 2-е, испр. и доп. – М., 1975.

3. Граф А.Е. Словарь немецких и русских пословиц и поговорок. – СПб.: Лань, 1997. – 228 с.

4. Жуков В.П. Словарь русских пословиц и поговорок. 5-изд. – М.: Рус. яз., 1993. – 536 с.

5. Кожемяко В.С., Подгорная Л.И. Русские пословицы и поговорки и их немецкие аналоги. – СПб.: Каро, 2000. – 192 с.

6. Носенко И.А. Начала статистики для лингвистов. Учеб. пособие для студ. пед. интов по спец. «Иностранные языки» – М.: Высш. школа, 1981. – 157 с.

7. Фрумкина Р.М. Некоторые практические рекомендации по составлению частотных словарей // Русск. яз. в нац. школе, 1963, № 5. – С. 73 – 77.

8. Фрумкина Р.М. Статистические методы изучения лексики. – М., 1964. – 115 с.

9. Частотный словарь русского языка / Под редакцией Л.Н.Засориной. – М., 1977. – 935 с.

Гродненский государственный
университет имени Я. Купалы

Поступило 23.02.06