КРИТИКА
И СЕМИОТИКА

Корпуса со специализированной разметкой для изучения статистики концептов

Линь Цзиньфэн [1], Д. М. Семёнова [2], С. Л. Пущин [3]
Т. Г. Петров [4], М. Н. Бабарико [5], С. В. Чебанов [6]

[1] ЛАНЬЧЖОУСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
[2] ООО «ИНТЕЛЛИДЖЕР»
[3] ТИПОГРАФИЯ КСИ-ПРИНТ
[4] ООО «СОКОЛОВ»
[5] САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ
[6] САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Аннотация. Изучение статистики концептов предполагает работу с размеченными корпусами. В принципе, такая разметка может быть только ручной на основе экспертных оценок с привлечением нескольких экспертов. Однако в ряде случаев такая возможность исключена, и разметка делается одним разметчиком – автором исследования. Экспликация принципов разметки и воспроизводимые количественные закономерности (покрытие 80 % использования концептов 7 ± 2 из них), полученные на материале русского, китайского, французского, английского языков семью разметчиками (6 русскими и 1 китаянкой, 6 девушками и 1 юношей), дают основание считать такую разметку удовлетворительной.

Линь Цзиньфэн, Семёнова Д. М., Пущин С. Л., Петров Т. Г., Бабарико М. Н., Чебанов С. В. Корпуса со специализированной разметкой для изучения статистики концептов // Критика и семиотика. 2020. № 2. С. 87–113.

ISSN 2307-1737. Критика и семиотика. 2020. № 2

Ключевые слова: концепт, распределение концептов, квантитативная концептология, ручная разметка текста, соотношение Парето, магическое число Миллера.

УДК 81’32 + 81’44

DOI 10.25205/2307-1737-2020-2-87-113

Контактная информация: Линь Цзиньфэн, кандидат филологических наук, преподаватель Ланьчжоуского политехнического университета (Lanzhou, China, linjinfeng1990@163.com)

Семёнова Дарья Михайловна, директор по развитию ООО «Интеллиджер» (Санкт-Петербург, Россия, dasha.glc@gmail.com)

Пущин Сергей Львович, генеральный директор, ООО «Типография КСИ-Принт» (Санкт-Петербург, Россия, z1q813@mail.ru)

Петров Томас Георгиевич, доктор геолого-минералогических наук, профессор-консультант, ООО «Соколов» (Санкт-Петербург, Россия, tomas_petrov@rambler.ru)

Бабарико Максим Николаевич, аспирант, Санкт-Петербургский государственный экономический университет (Санкт-Петербург, Россия, maxbabaro@gmail.com)

Чебанов Сергей Викторович, доктор филологических наук, профессор, кафедра математической лингвистики филологического факультета СПбГУ (Санкт-Петербург, Россия, s.chebanov@gmail.com)

Вводные замечания

Корпусный подход к исследованию концептов открывает принципиально новые возможности, но и порождает новые затруднения. Первые связаны с тем, что облегчается доступ к большим массивам фактического материала. Вместе с тем именно величина этого материала оказывается основным затруднением, возникающим в случае привлечения корпусного материала. Приемлемым выходом является привлечение к работе с корпусами компьютерных технологий. Однако большая часть последних предполагает возможность формализованного представления материала и формальных методов работы с ним. Это обстоятельство сужает возможности корпусных исследований семантики, ограничивая последние грамматической семантикой и в малой степени охватывая сферу семантики лексической.

Если же речь идет о лексической семантике, то очевидным путем для исследования концептов является разметка корпуса несколькими исследователями, выступающими в роли экспертов, и сопоставление результатов такой разметки.

1. Ручная разметка, произведенная одним разметчиком, в исследованиях по лингвосоциологии

Было проведено несколько исследований, в которых привлечение нескольких экспертов оказалось невозможным, и в качестве эксперта выступал сам автор исследования в сотрудничестве с С. В. Чебановым как руководителем.

Первый цикл таких исследований связан с изучением социальных институтов, представленных в текстах, которые маркированы как эталонные носители представлений о социальных институтах и являются типичными концептами. Частота упоминания таких институтов как концептов определялась по следующим источникам: «Народные русские сказки» А. Н. Афанасьева [Чернышова, 2008]; том 1 «О лицах» кодекса Наполеона [Ляпунова, 2010]; «Соборяне» Н. С. Лескова [Кириллова, 2008; Кириллова, Чернявский, 2009]; «История одного города» и «Сказки» М. Е. Салтыкова-Щедрина, «Жизнь и необычайные приключения солдата Ивана Чонкина» и «Москва 2042» В. Н. Войновича [Смирнова, 2008]; письменные работы младших школьников 1990-х и 2000-х гг. [Курочкина, 2008], что было обобщено в статье С. В. Чебанова [2012]. Во всех этих текстах концепты описываются довольно большим числом токенов – лексических единиц, словосочетаний и описательных конструкций.

В указанных случаях работа начиналась с того, что некоторый текст подвергался специализированной семантической разметке – в нем выделялись обозначения социальных институтов. Такой размеченный текст (или их совокупность) выступает с этого момента как корпус, который может изучаться с теми или иными целями.

Проблема при этом заключается в том, что при разметке таких текстов речь идет в первую очередь об институтах дальнего порядка (эта цель исследования была сформулирована в рамках реализации некоторого большого лингвосоциологического проекта, осуществлявшегося Институтом национальной модели экономики под руководством В. А. Найшуля [Найшуль, 2006]), которые в русской культуре не сформированы [Найшуль, Чебанов, 2010]. Это проявляется в том, что русские люди не в состоянии осуществлять адекватные социальные действия безотносительно к личным отношениям с представителями этих институтов. Так, обычный человек не в состоянии осознать, что если участковый полицейский, врач в поликлинике, водопроводчик в компании, обслуживающей дом, в котором живет этот человек, является его соседом или родственником, то функциональные отношения с ними должны строиться безотносительно к соседству или родству. Невозможность осознания этого является источником неизбывной русской коррупции, прежде всего повседневной, низовой. Но именно такое положение дел приводит к тому, что обычный носитель русского языка не в состоянии распознавать в сказках, романах, рассказах, эпосе, стихах и т. д. социальные институты, причем институты дальнего

социального порядка в первую очередь, а значит, он не способен и производить соответствующую семантическую разметку.

Не является выходом в этом случае и привлечение в качестве экспертов профессионалов-социологов, поскольку проблемы институтов дальнего порядка в отечественной социологии не сформированы (как и сами институты), и выделение таких институтов в обсуждаемых текстах будет в первую очередь определяться не этими текстами, а социологическим направлением, к которому принадлежит исследователь, выступающий в качестве эксперта (см. об этом, например, [Касьянова, 2003; Чеснокова, 2010; Щепаньский, 1969]). Полученные разметки будут в этом случае, скорее всего, несопоставимыми.

В силу указанных обстоятельств был выбран вариант работы, при котором авторы работ в течение нескольких месяцев изучали социологическую литературу по плану, разработанному руководителем (С. В. Чебановым), получали от него общие указания по разметке текстов и переходили к их разметке, имея возможность в любой момент консультироваться с руководителем (а иногда и с В. А. Найшулем) и иметь с его стороны выборочную проверку проделанной разметки.

Результаты, полученные таким образом, даже если считать их полученными не вполне бесспорными методами, оказываются интересными хотя бы в силу того, что не существует никаких других данных, позволяющих дать альтернативную количественную оценку. При этом они выглядят правдоподобными и допускающими содержательную интерпретацию, хотя часть из них представляется неожиданной.

2. Распределения концетов описания жестов

Несмотря на то что были получены количественные данные, в этом блоке исследований не ставилась задача характеристики распределений величин. Последние получены в работе Д. М. Семёновой, посвященной распределению частот описаний жестов в первом томе романа Л. Н. Толстого «Война и мир» [Семёнова, 2012; Семёнова, Чебанов, 2012]. Методика разметки текста была принципиально такой же, как и в случае социальных институтов, однако авторами разработана специфическая система тегов, для которых вводились отношения зависимости (нечто при условии, что…). При этом фиксировались не только количественные характеристики описаний отдельных жестов, но и статистические распределения таких характеристик (рис. 1). Вид распределений 1–6 был несколько неожиданным, но их изучение не входило в задачи указанного исследования.

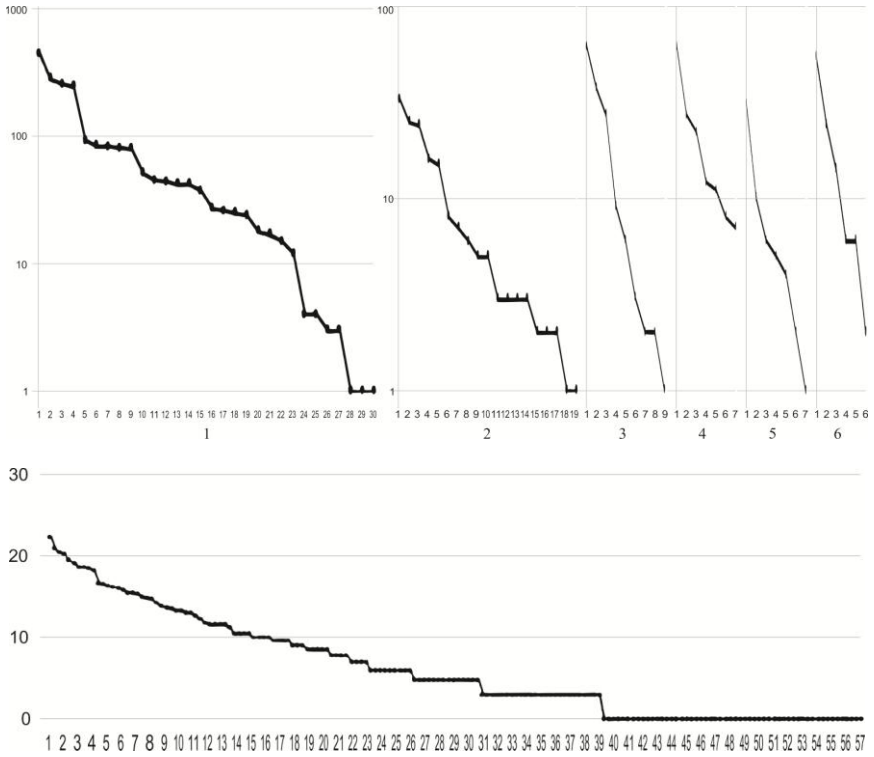


Рис. 1. Сопоставление распределения лексических средств, описывающих жесты героев (*1* – распределение элементов описания невербальной коммуникации; *2* – распределение элементов описания жестов князя Андрея Болконского; *3* – распределение ключевых элементов по числу описывающих характеристик; *4* – распределение ключевых элементов по группам; *5* – распределение ключевых элементов, охарактеризованных хотя бы одним признаком; *6* – соотношение двух элементов характеристик жестового знака), с типичным *H*-распределением жестов по героям в 1-м томе романа (всё в полулогарифмических координатах). Везде по оси абсцисс – ранг токена, по оси ординат – частота токена в тексте

Fig. 1. Comparison of the distribution of lexical means that describe geatures of characters (*1* – the distribution of elements about the description of non-verbal communication, *2* – the distribution of elements about the description of gestures of prince Andrei Bolkonsky, *3* – the distribution of key elements in the number of descriptive characteristics, *4* – the distribution of key elements in groups, *5* – the distribution of key elements, characterized by at least one feature, *6* – the relation of two elements of characteristics of the sign), with the typical *H*-distribution of gestures on the characters in the first volume of the novel (all in semi-logarithmic coordinates). Everywhere: on the abscissa axis – the rank of the token, on the ordinate axis – the frequency of the token in the text

3. Разметка пословиц для изучения составности человека

Распределение концептов было основным объектом исследования Линь Цзиньфэн [2018], посвященного концептам *ТЕЛО, ДУША, ДУХ* в русской и китайской пословичных картинах мира.

В этом случае также невозможно было проведение исследования с привлечением для разметки нескольких экспертов. Дело в том, что отсутствует единая антропологическая концепция, которая могла бы быть положена в основу сопоставления результатов разметки пословиц разными экспертами и оценки качества этой разметки. Такое положение определяется тем, что в европейской традиции есть две антропологические традиции – дихотомия, признающая двусоставность человека (тело и дух), и трихотомия, различающая в человеке тело, душу и дух, на базе которых формируются разные конфигурации концептов [Линь Цзиньфэн, Чебанов, 2018]. Эксперты, которые являются активными сторонниками каждой из точек зрения, будут размечать один и тот же текст совершенно по-разному. К тому же таких экспертов практически не найти. Обучить каждой из концепций какие-то группы испытуемых, которые будут использоваться далее как эксперты, – не реалистично. Поэтому разметка производилась автором (Линь Цзиньфэн), который на протяжении примерно двух с половиной лет занимался изучением этой проблемы, после чего выбрал трихотомию как концепцию, на базе которой производится анализ, и приступил к разметке корпусов пословиц, консультируясь со своим руководителем (С. В. Чебановым). Методика разметки подробно описана во второй главе ее диссертации [Линь Цзиньфэн, 2018].

3.1. Источники исследованного материала

Материалом для исследования выступали эталонные корпуса пословиц: для русских пословиц собрание В. И. Даля «Пословицы русского народа» (1861–1862), а для китайских 中国谚语资料 – Собрание китайских пословиц Ланчжоуского института искусств (1961, 1962). В этих собраниях были выделены пословицы, в которых фигурируют реалии, соотносимые с концептами *ТЕЛО, ДУША, ДУХ* (табл. 1).

3.2. Распределения токенов, выражающих концепт

На основании проведенной ручной разметки был собран материал для дальнейшей статистической обработки, которая дала следующие результаты. Если для каждого из языков взять токены [1] (лексемы, словосочетания,

[1] Предлагаемое использование термина, обычное для английского языка, несколько отличается от принятого в русском, хотя трактовка токена как значимой

описательные конструкции, косвенные обозначения и т. д.), представляющие все три концепта (включая и описания их частей), то (после исключения первых наиболее высокочастотных классов – ср. [Фуфаев, 2009]) получается обычное ранговое распределение гиперболического вида (рис. 2) или в полулогарифмических координатах (рис. 3). Во втором случае имеется практически линейная зависимость, нарушаемая на первых рангах сверхвысокочастотными классами (исключенными из графиков рис. 2, 3) и ступеньками в области больших рангов за счет встречаемости нескольких или многих равночастотных классов.

Таблица 1

Число пословиц с концептами *ТЕЛО, ДУША, ДУХ*

Table 1

Number of proverbs with concepts *BODY, SOUL, SPIRIT*

Количество пословиц	Концепт			Всего
	ТЕЛО	*ДУША*	*ДУХ*	
В Собрании Даля				
Абсолютное	7 611	5 935	1 500	11 584 *
% от общего количества в собрании (31 351)	24,28	18,93	4,78	36,95
% от общего количества с концептами (11 584)	65,70	51,23	12,95	100,00
В Собрании китайских пословиц				
Абсолютное	5 147	3 080	1 020	7 252
% от общего количества в собрании (32 576)	15,80	9,45	3,13	22,26
% от общего количества с концептами (7 252)	70,97	42,47	14,07	100,00

* 7 611 + 5 935 + 1 500 > 11 584; 24,28 + 18,93 + 4,78 > 36,95; 65,70 + 51,23 + + 12,95 > 100,00 и аналогично в китайских из-за того, что одна пословица может содержать более одного концепта.

единицы языка [Захаров, Богданова 2013, с. 140] или трактовка А. М. Карамнова [2014, с. 83] позволяют использовать его в таком смысле.

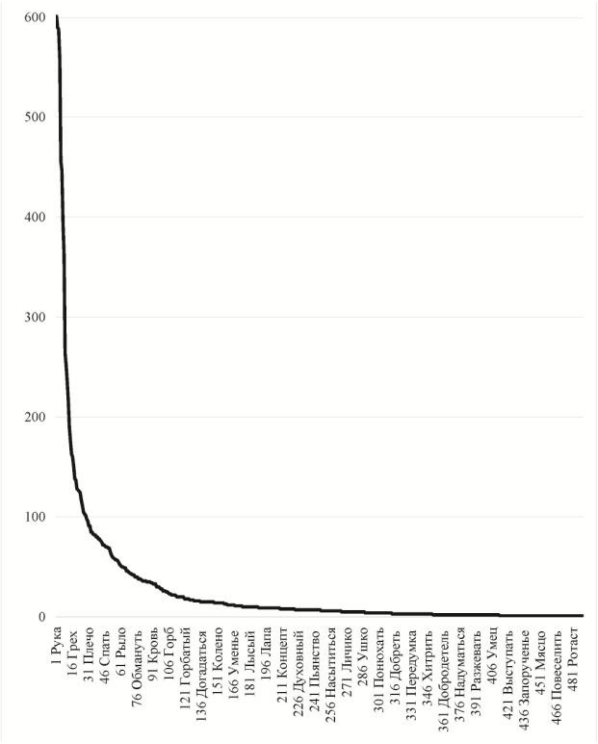


Рис. 2. Частотно-алфавитное распределение лексем описания составности человека в русских пословицах (за исключением первых четырех наиболее частотных)
Fig. 2. Frequency-alphabetical distribution of lexemes describing the composition of human in Russian proverbs (except for the first four most frequent)

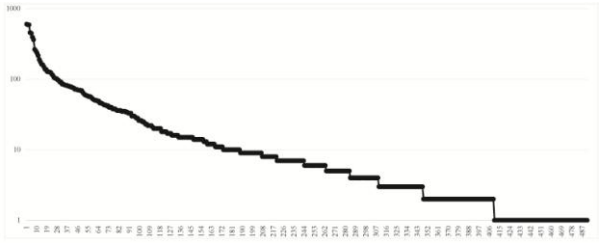


Рис. 3. Частотно-алфавитное распределение лексем описания составности человека в русских пословицах (за исключением первых четырех наиболее частотных в полулогарифмических координатах)
Fig. 3. Frequency-alphabetical distribution of tokens describing the composition of human in Russian proverbs (except for the first four most frequent in semi-logarithmic coordinates)

3.3. Распределение концептов, представляющих составность человека

На основе анализа семантики токенов и содержащих их пословиц были выделены концепты 1-го уровня, передающие составность человека в пословичных картинах мира, и вычислены их частоты как суммы частот, передающих их токенов (табл. 2).

Таблица 2

Концепты русских и китайских пословиц (первый уровень)

Table 2

Concepts of Russian and Chinese peoverbs (first level)

R	Русские пословицы			Китайские пословицы		
	концепт	количество		концепт	количество	
		ед.	%		ед.	%
1	*ДУША* (неспец.)	5 152	19,82	*РОТ*	1 377	13,96
2	*ТЕЛО*	4 058	15,61	*СЕРДЦЕ*	1 318	13,36
3	*ГОЛОВНАЯ ДУША*	3 235	12,44	*ГЛАЗ*	825	8,36
4	*ДУХ*	2 356	9,06	*СЕРДЕЧНАЯ ДУША*	732	7,42
5	*РОТ*	2 028	7,80	*ЖИВОТ*	570	5,78
6	*СЕРДЕЧНАЯ ДУША*	1 797	6,91	*КИСТЬ*	558	5,66
7	*ГЛАЗ*	1 247	4,80	*ТЕЛО*	503	5,10
8	*ГОЛОВА* (неспец.)	851	3,27	*ГУБА*	410	4,16
9	*ТЕЛО + ДУША + ДУХ*	715	2,75	*СТОПА*	406	4,12
10	*<КИСТЬ>*	601	2,31	*ГОЛОВА* (неспец.)	405	4,11
11	*ЛИЦО* (неспец.)	315	1,21	*ЛИЦО* (неспец.)	314	3,18
12	*УХО*	306	1,18	*ДУХ*	295	2,99
13	*ДУША + ДУХ*	300	1,15	*ГОЛОВНАЯ ДУША*	286	2,90
14	*<СТОПА>*	264	1,02	*УХО*	255	2,59
15	*НОГА*	216	0,83	*НОГА*	196	1,99
16	*РУКА*	193	0,74	*МЕРТВЕЦ*	190	1,93
17	*ЖЕЛУДОК*	190	0,73	*РУКА*	173	1,75
18	*ЖИВОТ*	164	0,63	*ПОЯСНИЦА*	104	1,05

Окончание табл. 2

R	Русские пословицы			Китайские пословицы		
	концепт	количество		концепт	количество	
		ед.	%		ед.	%
19	*ЗУБ*	156	0,60	*ЯЗЫК*	97	0,98
20	*ГОРЛО*	141	0,54	*КОЖА*	87	0,88
21	*СПИНА*	141	0,54	*ПЕЧЁНОЧНАЯ ДУША*	84	0,85
22	*НОС*	138	0,53	*ЗУБ*	76	0,77
23	*СЕРДЦЕ*	127	0,49	*ГОРЛО*	65	0,66
24	*ПОКОЙНИК*	126	0,48	*СПИНА*	61	0,62
25	*ПАЛЕЦ*	99	0,38	*ЖЁЛЧНЫЙ ПУЗЫРЬ*	59	0,60
26	*ПЛЕЧО*	93	0,36	*КОСТЬ*	55	0,56
27	*ГУБА*	90	0,35	*НОС*	54	0,55
28	*БОК*	86	0,33	*КУЛАК*	52	0,53
29	*КОЖА*	86	0,33	*СЛЕЗА*	47	0,48
30	*СЛЕЗА*	86	0,33	*БРОВЬ*	44	0,45
31	*ЯЗЫК*	84	0,32	*ПЛЕЧО*	38	0,39
32	*ГРУДЬ*	71	0,27	*ГРУДЬ*	32	0,32
33	*КОСТЬ*	70	0,27	*ПЕЧЕНЬ*	32	0,32
34	*КУЛАК*	65	0,25	КРОВЬ	30	0,30
35	*НОГОТЬ*	45	0,17	*ДУША* (неспец.)	12	0,12
36	*ЛОБ*	43	0,17	*ЩЕКА*	12	0,12
37	*ГОЛЫЙ*	42	0,16	*ЛЁГКОЕ*	9	0,09
38	*КРОВЬ*	33	0,13			
39	*ЩЕКА*	26	0,10			
40	*БРОВЬ*	24	0,09			
41	*ЛОКОТЬ*	24	0,09			
42	*ПЕЧЁНОЧНАЯ ДУША*	20	0,08			
43	*ПЯТКА*	20	0,08			
44	*КУКИШ*	18	0,07			
45	*ТИТЬКА*	15	0,06			
46	*КОЛЕНО*	14	0,05			
47	*ПУП*	11	0,04			
48	*ПЕЧЕНЬ*	8	0,03			
49	*<МАТКА>*	7	0,03			

Далее на основании отношения «часть – целое» были выделены концепты 2-го и последующих уровней (всего 5 для обоих языков, хотя условие равенства при проведении исследования изначально не задавалось). Их частоты вычислялись как суммы частот входящих в них концептов более низкого уровня (табл. 3).

Таблица 3

Число концептов разных уровней
и их число, приходящееся на 80 % покрытия

Table 3

Number of concepts in different levels
and their number accounted for 80% of the coverage

Уровень концепта	Русские пословицы		Китайские пословицы	
	количество	80 % покрытия	количество	80 % покрытия
1	49	8	37	12
2	36	8	27	9
3	24	6	21	7
4	18	6	15	5
5	5	2	3	1

При этом оказывается, что распределения частот концептов, передающих составность человека, столь резко неравночисленные и в ранговой форме настолько резко убывающие, что их удобно представлять в полулогарифмических координатах (рис. 4).

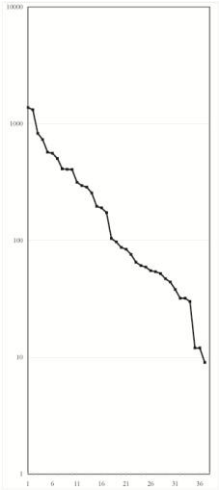


Рис. 4. Частоты концептов, передающих составность человека, 1-го уровня в китайских пословицах (в полулогарифмических координатах)
Fig. 4. Frequency of concepts that transmit human composition, first level in chinese proverbs (in semi-logarithmic coordinates)

Характер этих распределений оказывается однотипным для распределений частот концептов 1÷5 уровней русских и китайских пословиц (рис. 5). При этом чем выше уровень концептов, тем их меньше и тем круче оказывается падение их частот.

В связи с таким резким падением частот концептов была изучена динамика накопления их частот. Также обнаружилась совершенно однотипная картина, показанная на рис. 6 (концепты третьего уровня выбраны в связи с тем, что их число сравнительно невелико, а картина динамики совершенно такая же, как и для концептов других уровней русских и китайских пословиц).

Примечательно и то, что соотношение Парето 80 : 20, выступающее в данном случае как 80 % покрытия частот употребления всех концептов, достигается для концептов 1÷4 уровней русских и китайских пословиц за счет 5÷9 концептов и слабо зависит от общего числа концептов (15÷49; см. табл. 3). Это в точности соответствует магическому числу Миллера 7 ± 2 компонентов, которыми оптимально манипулирует оперативная память [Миллер, 2010]. Единственным исключением являются 12 для китайских концептов первого уровня, что может быть связано с особенностями строя китайского языка. Для 5-го уровня число концептов так мало, что данное отношение теряет смысл.

4. Распределение числовых концептов пословиц

Совершенно аналогичная картина была получена и для концептов чисел в пословицах собраний В. И. Даля [1862] и В. М. Мокиенко с соавторами [Мокиенко и др., 2010], в данных М. Н. Бабарико [Бабарико, Чебанов, 2014; 2015]), в собрании китайских пословиц [Babariko et al., 2016] и в новых данных настоящей работы (рис. 7). При этом во всех трех случаях 80 % покрытия обеспечивается 7 числами.

5. Распределение токенов, выражающих концепт

Другим направлением квантитативной концептологии может быть изучение распределений не концептов, а частот токенов, с помощью которых выражаются концепты.

5.1. Пойнтер-точка и характеристические лексемы концепта

Один из вариантов таких исследований был намечен Б. И. Кудриным [2007] в связи с его представлением о пойнтер-точке. Она определяется следующим образом.

Видовым (спектровым) представлением *H*-распределения токенов является функция $\Omega(x)=W_0/x^{1+\alpha}$, где $\Omega(x)$ – количество токенов с одинаковым количеством употреблений, x – количество употреблений каждого токена, α – характеристический показатель распределения, а за W_0 принимается численность самого частого токена.

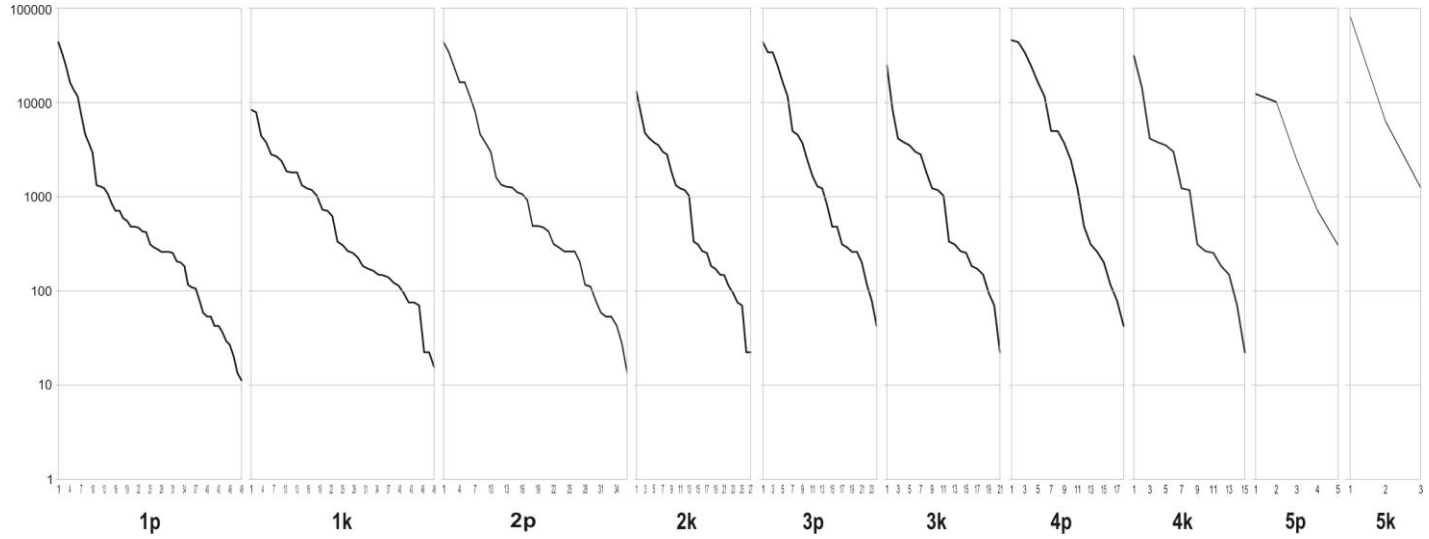


Рис. 5. Распределение частот концептов, передающих составность человека, 1÷5-го уровней в русских (p) и китайских (k) пословицах (в полулогарифмических координатах; масштаб по оси абсцисс точно не выдержан из-за несоизмеримости числа концептов разных уровней)

Fig. 5. Distribution of frequencies of concepts representing composition of human being 1÷5 levels in Russian (p) and Chines (k) proverbs (in semi-logarithmic coordinates; the scale along the abscissa axis is not exactly sustained due to the incommensurability of the number of concepts of different levels)

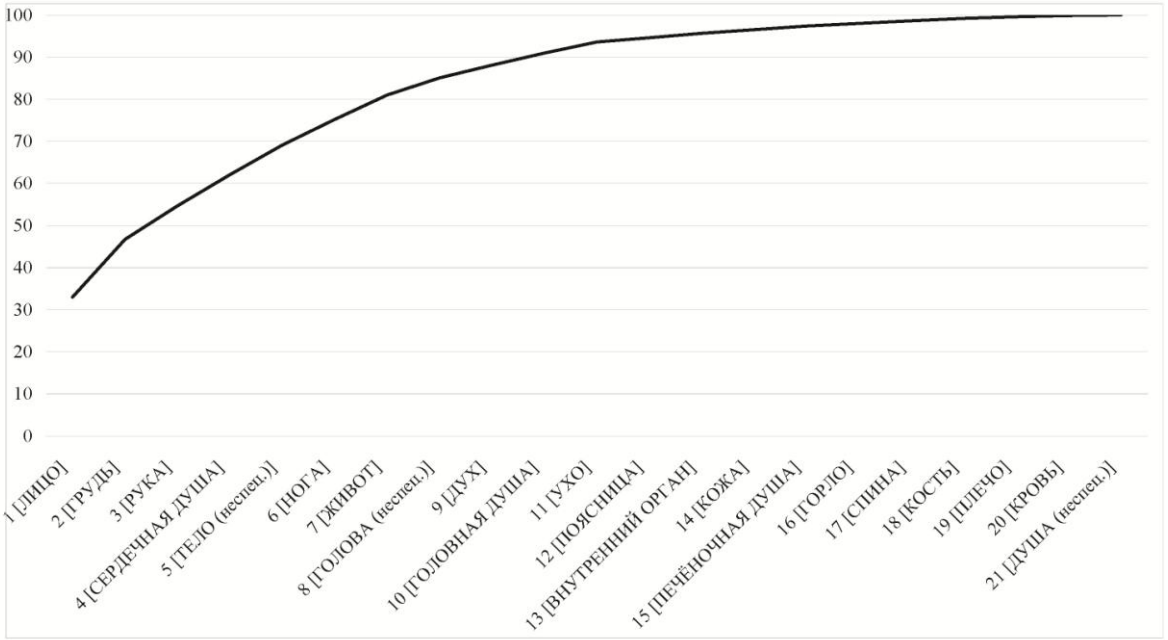


Рис. 6. Накопленные частоты концептов третьего уровня китайских пословиц
Fig. 6. Accumulated frequencies of third-level concepts of Chinese proverbs

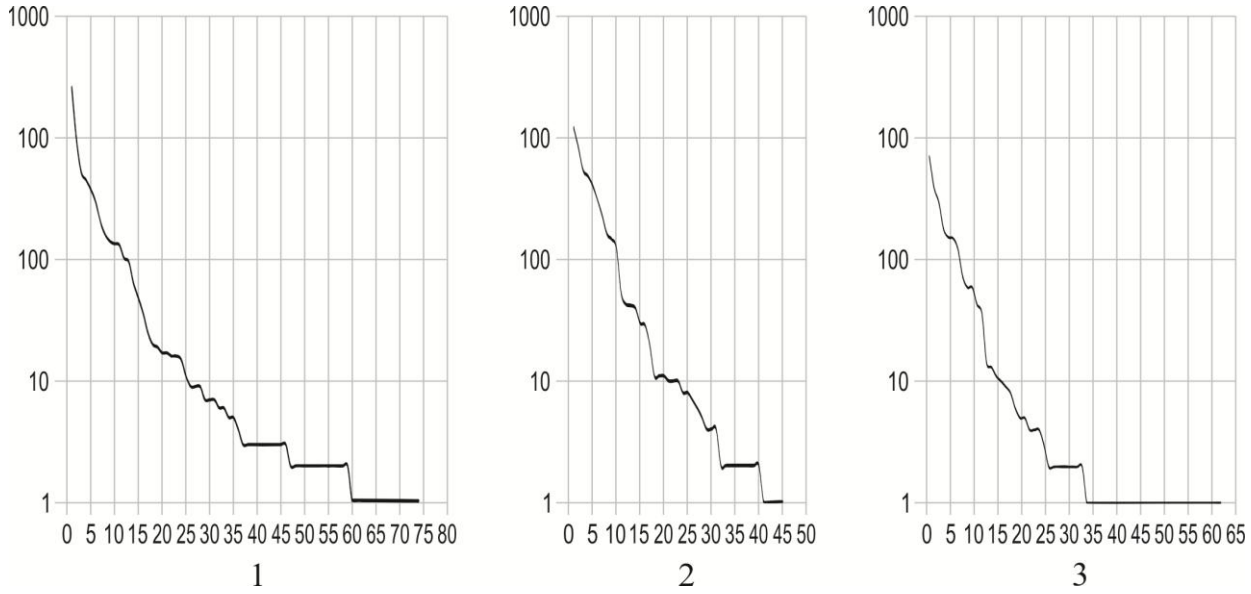


Рис. 7. Распределение концептов чисел:
1 – собрание китайских пословиц; *2* – собрание пословиц В. И. Даля; *3* – собрание пословиц В. М. Мокиенко и др.
Fig. 7. The distribution of numbers' concepts:
1 – in Chinese proverbs; *2* – in V. I. Dal's collection of proverbs; *3* – in V. M. Mokienko et al.'s collection of proverbs

Пойнтер-точка ℜ такая, что «гипербола делится точкой ℜ на две ветви: слева i = 1, 2, …, ℜ – неоднородные касты <классы знакотипов, в данном случае лексем, с одинаковой частотой>, где каждая каста представлена множеством видов <множеством лексем>; справа i = ℜ +1, ℜ +2, ..., ℜ + + K – однородные <содержащие только одну лексему> касты … (i соответствует числу особей этого вида)» [Кудрин, 2007 с. 29].

Исследование Е. Б. Кудриной показывает, что при изучении частоты упоминаний героев в романе М. А. Булгакова «Мастер и Маргарита» к пойнтер-точке ℜ =34 примыкают Левий Матвей, Гелла, Н. И. Босой, Варенуха, Римский, Стёпа Лиходеев и Га-Ноцри, которые и отражают булгаковскую специфику повествования и его отличие от «Фауста» Гёте (см. [Кудрин, 2007, с. 31]). В связи с этим был предпринят поиск пойнтер-точки для распределения токенов, описывающих составность человека в русских и китайских пословицах. Для этого был использован способ расчета пойнтер точки по методике С. Л. Пущина [2014], суть которой заключается в следующем.

На основании полученного специальным методом подбора коэффициента β = 1 + α программой, составленной С. Л. Пущиным, рассчитывается идеальная численность класса, соответствующая касте ℜ, $N_{ℜ}$, после чего в эмпирическом распределении ищется лексема с частотой, наиболее близкой к $N_{ℜ}$, которая и принимается за пойнтер-точку ℜ.

В результате было обнаружено, что пойнтер-точка для распределения русских лексем соответствует частоте 186 (ближайшие лексемы Воля – 190 и Сердце – 175), а для китайских – 92 (ближайшие лексемы 听 Слышать – 96, 腿 Нога от пояса до стопы – 87). Среди русских лексем вблизи пойнтер-точки оказываются лексемы *добро*, *видеться*, *добрый*, *воля*, *сердце*, *чёрт*, *грех*, характеризующие *СЕРДЕЧНУЮ ДУШУ*, *ГОЛОВНУЮ ДУШУ*, *ДУХ*, а через *видеть* и *ТЕЛО*, что вполне соответствует образу русского человека как живущего душой и духом. Среди лексем китайских пословиц вблизи пойнтер-точки находятся *饱 сыт*, *死 умереть*, *饥 голод*, *讲* (книжн.) *говорить*, *食* (книжн.) *еда*, *腰 поясница*, которые передают потребность в еде как важнейшую потребность *ТЕЛА*, отмечая и такую важную для китайской культуры часть *ТЕЛА*, как *ПОЯСНИЦА* (которая присутствует на 3-х из 5-ти уровней иерархии концептов), включая только две лексемы, обозначающие действия *ГОЛОВНОЙ ДУШИ – слышать* и *говорить*. Картина оказывается весьма контрастной и соответствующей клишированным образам представителей двух народов, что позволяет рассматривать нахождение пойнтер-точки как перспективный способ нахождения токенов, претендующих на представление ядра концепта

(в данном случае концепта *ЧЕЛОВЕК* в русской и китайской пословичных картинах мира).

При этом по материалам Д. М. Семёновой оказалось возможным для концепта *ВЗГЛЯД* рассмотреть распределение концептов его вариантов (т. е. средств передачи концепта *ВЗГЛЯД*) и рассчитать для него пойнтер-точку. Оказалось, что ей соответствует концепт *СМОТРЕТЬ* с частотой 9 в окружении *ОПУСТИТЬ ГЛАЗА* и *ВЗГЛЯД*.

5.2. Энтропийно-анэнтропийный анализ средств выражения концептов

Другим способом количественного анализа токенов, выражающих концепт, является энтропийно-анэнтропийный анализ (метод *RHA* Т. Г. Петрова [2008]). Анализ заключается в том, что после получения ранговой формулы *R* (которой в данном случае является частотный словарь токенов, с помощью которых выражается тот или иной концепт или их группа) осуществляется вычисление информационной энтропии Шеннона ($H = -\Sigma p_i * \ln p_i$, где p_i – нормированная к 1 частота *i*-го токена [Петров, Фарафонова, 2005, с. 48]), характеризующей равномерность распределения токенов, и анэнтропии ($A = -[(\Sigma \ln p_i)/n] - \ln(n)$, где n – число токенов, представляющих концепт [Там же, с. 61]), введенной Т. Г. Петровым для характеристики неравномерности вкладов компонентов (в данном случае токенов) в их распределение, на основании сопоставления которых у изучаемых объектов делаются содержательные выводы. При этом *H* и *A* могут рассчитываться для полных (что позволяет полнее охарактеризовать их индивидуальность) или усеченных (обрубленных) составов объектов (с целью их сопоставления вне зависимости от природы объектов и способов их изучения).

Для концептов, представляющих составность человека, были вычислены *H* и *A* для токенов, представляющих концепты *ТЕЛО*, *ДУША*, *ДУХ* и их совокупности в русских и китайских пословицах (табл. 4).

Таблица 4

Число токенов, представляющих концепт

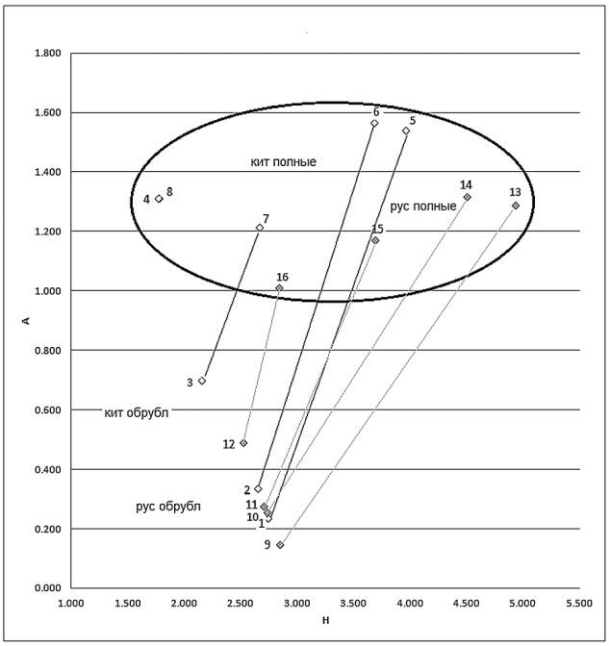
Table 4

Number of tokens representing the concept

Концепт	Число токенов, представляющих концепт	
	китайские пословицы	русские пословицы
ТЕЛО + ДУША + ДУХ	246	490
ТЕЛО	182	336
ДУША	60	119
ДУХ	20	43

При этом расчеты были произведены по полным и усеченным (в последнем случае по 20 самым частым токенам – минимальному числу токенов, выражающих концепт, в данном случае концепт *ДУХ* китайских пословиц) словарям токенов.

В итоге получены в целом ожидаемые результаты (рис. 8), которые, тем не менее, ставят некоторые частные вопросы.



	H	A		
1	2.736	0.237	тело+душа+дух	Кит
2	2.653	0.337	тело	Кит
3	2.150	0.702	душа	Кит
4	1.772	1.314	дух	Кит
5	3.960	1.542	тело+душа+дух	Кит
6	3.684	1.566	тело	Кит
7	2.663	1.216	душа	Кит
8	1.772	1.314	дух	Кит
9	2.847	0.149	тело+душа+дух	Рус
10	2.730	0.258	тело	Рус
11	2.702	0.277	душа	Рус
12	2.524	0.491	дух	Рус
13	4.928	1.291	тело+душа+дух	Рус
14	4.505	1.319	тело	Рус
15	3.688	1.174	душа	Рус
16	2.838	1.013	дух	Рус

Рис. 8. Энтропийно-анэнтропийные характеристики наборов токенов, реализующих концепты. Точки 4 и 8 совпадают в связи с тем, что число токенов, реализующих концепт *ДУХ* в китайских пословицах, принято за объем усеченного словаря

Fig. 8. Entropy-anentropy characteristics of tokens sets that implement concepts. Point 4 and 8 coincide due to the fact that the number of tokens implementing the concept *SPIRIT* in Chinese proverbs is taken as the volume of the truncated dictionary

Изменение величины энтропии *H* совершенно однотипно и понятно для полных и усеченных словарей русских и китайских пословиц. Для каждого из четырех сопоставимых комплектов данных *H* является минимальной для словарей, представляющих концепт *ДУХ*, которые являются самыми короткими во всех случаях. Следующие по величине – энтропии для токенов концепта *ДУША*, более детально представленного в русских и китайских пословицах, а еще больше для концепта *ТЕЛО*, наиболее подробного представленного в пословицах. Естественно, *H* для токенов, передающих совокупность концептов *ТЕЛО* + *ДУША* + *ДУХ*, является максимальной.

Сравнение этих четырех комплектов друг с другом дает более сложную картину: для полных и усечённых составов все величины *H* для русских пословиц сдвинуты в сторону больших величин на два шага, так что *H* концепта *ТЕЛО* китайских пословиц сопоставима с таковой концепта *ДУХ* русских, что свидетельствует о большей дифференциации представлений о человеке в русских пословицах. При этом основной вклад в *H* совокупности концептов *ТЕЛО* + *ДУША* + *ДУХ* вносят токены, передающие концепт *ТЕЛО*, причем этот вклад заметно больше для китайских пословиц, в то время как разницы значений *H* для словарей *ДУША – ТЕЛО* и *ДУХ – ДУША* примерно одинаковы. При этом, как это следует из самой логики расчетов, начинаясь с одной и той же минимальной величины, для полных словарей размах варьирования *H* оказывается примерно в два раза больше, чем для усеченных.

Таким образом, использование *H* для оценки разнообразия словарей качественно ничего нового не приносит, однако позволяет давать порядковым оценкам (больше – меньше) количественную оценку.

Ситуация с анэнтропией оказывается иной – характер ее изменений для полных и усеченных словарей оказывается совершенно разным.

Для усеченных словарей анэнтропия *A* уменьшается с увеличением словаря, т. е. с ростом энтропии *H*, что вполне понятно, так как *A* является характеристикой своеобразия словаря, которое связано прежде всего с редкими токенами, доля которых при фиксированном количестве токенов, принимаемых в расчет, уменьшается по мере увеличения полного словаря. Неожиданным оказалось то, что все точки, за исключением одной (12 – концепт *ДУХ* русских пословиц) лежат на одной прямой, причины чего требуют прояснения.

Для полных словарей значения величин анэнтропии *A* в несколько раз больше, чем для усеченных, а характер изменения совершенно другой – есть тенденция к ее росту с увеличением объема словаря и соответственно ростом энтропии *H*. Эта тенденция вполне понятна, так как с увеличением словаря может увеличиваться и число токенов с низкой частотой. Поэтому объяснимо увеличение *A* при переходе от словаря токенов, выражающих концепт *ДУХ*, к выражающим концепт *ДУША*, а далее к выражающим

концепт *ТЕЛО*. Эта закономерность выражена для концептов русских пословиц (точки 16, 15, 14), но нарушается для китайских за счет понижения *A* при переходе к токенам, выражающим концепт *ДУША*, что связано, по-видимому, с тем, что для его выражения используется большое число высокочастотных специфических токенов. Однако при переходе от токенов, выражающих концепт *ТЕЛО*, к токенам совокупности концептов *ТЕЛО + ДУША + ДУХ* происходит некоторое падение *A* из-за того, что одни и те же токены могут использоваться в зависимости от контекста для передачи разных концептов. Поэтому в результате объединения словарей может поменяться статистический статус токенов. Такой эффект вполне объясним, хотя не был ожидаем.

6. Распределение Линь Цзиньфэн

Проведенное исследование дает основание говорить о том, что обнаружен новый статистический объект со своеобразным устойчивым набором характеристик. Его особенности таковы.

Имеется некоторый текст (например, собрания русских и китайских пословиц, 1-й том «Войны и мира», Гражданский кодекс Франции и т. д.), в котором выделен набор однотипных концептов, скажем, *ТЕЛО, ДУША, ДУХ*, однако их должно быть достаточно много – десять и более.

В таком случае распределение частот этих концептов в ранговой форме будет описываться резко убывающим распределением (показано для 12 массивов: составляющие человека в русских и китайских пословицах, числа в русских и китайских пословицах, описания жестов в «Войне и мире», социальные институты в «Соборянах» Лескова, два массива письменных работ младших школьников, Гражданский кодекс Франции, сатирическая проза Войновича, «Сказки» и «История одного города» Салтыкова-Щедрина, русские народные сказки в собрании Афанасьева), которое обладает следующими свойствами:

- миллеровское число 7 ± 2 концептов из этого набора покрывает 80 % употреблений концептов из этого набора (выполняется закон Парето 80 : 20; показано для 6 массивов – концептов составляющих человека в русских и китайских пословицах, описаний жестов в «Войне и мире», чисел в двух русских и одном китайском собраниях пословиц);
- как правило, частота самых редких концептов измеряется несколькими единицами – десятком в том случае, если эти концепты связаны с тематикой текста. Возможны ситуации более редкого или однократного употребления концепта (в тексте, тематика которого слабо связана с семантикой концепта; например, числа в пословицах), но они никогда не образуют длинного хвоста *hapax legomena*, как в случае *H*-распределения токенов (на которое похоже описываемое распределение при условии очень резкого падения *H*-распределения). Это дает основание квалифици-

ровать данные распределения как распределения с толстыми хвостами [Фуфаев, 1996; Anderson, 2006], которые обрезаны.

Распределение частот совокупности токенов, выражающих этот набор концептов в данном тексте, аппроксимируется *H*-распределением, вблизи пойнтер-точки *R* которого концентрируются токены, наиболее полно представляющие данный набор концептов (показано для 3 массивов – русских и китайских пословиц и описаний жестов в «Войне и мире»).

Для усеченных словарей токенов существует обратное соотношение энтропии *H* и анэнтропии *A*, отклонения от линейности которого, как можно полагать, связано с внутреннними свойствами текста (показано для 2 массивов – русских и китайских пословиц). Обнаруженная линейность соотношения между *H* и *A*, в отличие от многих иных совокупностей «родственных» данных на диаграммах *HA*, требует специального исследования для проверки неслучайности его возникновения при работе с токенами и выявления его причины, если таковая будет подтверждена.

Для полных словарей токенов характерно прямое отношение энтропии и анэнтропии, причем анэнтропия полнее отражает особенности текста (показано для 2 массивов – русских и китайских пословиц).

Обнаруженное распределение было предложено обозначать как распределение Линь Цзиньфэн [Линь Цзиньфэн и др., 2019].

7. Воспроизводимость характеристик распределения концептов как критерий качества разметки

Таким образом, на изученных с разной степенью подробности 12 массивах данных, которые можно рассматривать как размеченные корпуса, выявлены (что само по себе является новым) однотипные закономерности распределения концептов, перечисленные выше. Крайне сомнительно, что такой результат может считаться артефактом, тем более что получены и устойчивые количественные характеристики этих распределений.

Указанные обстоятельства, в свою очередь, могут рассматриваться как свидетельство того, что выявленные результаты основаны на данных приемлемого качества. Исходными же данными для всех видов статистической обработки были данные разметки текста. Для всех массивов текста такой разметкой была ручная разметка, при которой автор исследования выступал в качестве эксперта-разметчика.

Сходство качественных особенностей полученных результатов в сочетании с содержательной (каждый раз своей) интерпретируемостью полученных результатов, а также сходство их количественных характеристик может поэтому рассматриваться как свидетельство того, что произведенная ручная разметка, основанная на экспертных оценках одного разметчика, может рассматриваться как удовлетворительная.

Состоятельность сделанного утверждения обосновывается и тем, что разметка была осуществлена 7-ю разметчиками (в основном не знакомых

друг с другом; некоторые из них при этом явно привлекали помощников), из которых 6 девушек и 1 юноша, принадлежащих к разным языковым культурам (6 русских и 1 китаянка). Обработка данных велась разными методами (анализ частотных словарей лексем, частотных словарей концептов, энтропийно-анэнтропийный анализ, расчет пойнтер-точки), а расчеты производились разными лицами, в разных программах и на разном оборудовании. Несмотря на это, были получены воспроизводимые и в большинстве случаев концептуально соотносимые друг с другом результаты.

Всё это позволяет утверждать, что ручная семантическая разметка корпуса, основанная на экспертных оценках и осуществленная одним разметчиком, может рассматриваться в случае невозможности использования других методов как приемлемый, по крайней мере для изучения концептов, способ разметки.

Список литературы

Бабарико М. Н., Чебанов С. В. Арифмология русских пословиц и поговорок собрания В. И. Даля // Структурная и прикладная лингвистика. 2014. Вып. 10. С. 70–91.

Бабарико М. Н., Чебанов С. В. Русская паремиологическая арифмология XIX–XXI веков // Структурная и прикладная лингвистика. 2015. Вып. 11. С. 186–219.

Даль В. И. Пословицы русского народа. М.: В Университетской типографии, 1862. 883 с.

Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». СПб.: Изд-во СПбГУ, 2013. 148 с.

Карамнов А. С. Количественная оценка повторяемости и сложности лексики в корпусе учебника английского языка // Филологические науки. Вопросы теории и практики. 2014. № 6 (36), ч. 1. С. 82–86.

Касьянова К. О русском национальном характере. М.: Академический Проект; Екатеринбург: Деловая книга, 2003. 560 с.

Кириллова М. В. Концептуализация социальных институтов в «Соборянах» Н. С. Лескова: Выпускная квалификационная работа. СПб.: БГТУ, 2008. 105 с.

Кириллова М. В., Чернявский В. А. Концепты социальных институтов в произведении Н. С. Лескова «Соборяне» / Ин-т социологии РАН. Методологический семинар памяти Г. С. Батыгина, 2009. URL: http://www.isras.ru/files/File/Seminar/Seminar_Batygin/Kirillova_Chernyavsky.pdf

Кудрин Б. И. Мои семь отличий от Ципфа // Общая и прикладная ценология. 2007. № 4. С. 25–33.

Курочкина А. С. Социальные концепты в языковом творчестве детей младшего школьного возраста (диахронический анализ): Выпускная квалификационная работа. СПб.: БГТУ, 2008. 241 с.

Линь Цзиньфэн. Концепты [ТЕЛО], [ДУША], [ДУХ] в русской и китайской языковых картинах мира (антропологическая трихотомия в пословичной картине мира): Дис. … канд. филол. наук. СПб., 2018. Т. 1–2. 221 + 147 с.

Линь Цзиньфэн, Пущин С. Л., Петров Т. Г., Семёнова Д. М., Чебанов С. В. Усеченные ципфоподобные распределения и частотные словари концептов // Общая и прикладная ценология как приятие понимания фундаментальности природного закона видового разнообразия особей сообществ третьей научной картины мира материальной и идеальной реальностей. Практические исследования. Обобщающие материалы по общей и прикладной ценологии. Труды XXII Встречи-семинара ценологов (Москва, НИУ МЭИ, 16.11.2018) // Ценологические исследования. СПб.: КСИ-Принт, 2019. Вып. 59. С. 108–121.

Линь Цзиньфэн, Чебанов С. В. Формирование концептов [ТЕЛО], [ДУША], [ДУХ] в современной русской языковой картине мира // Вестник Тюмен. гос. ун-та. Гуманитарные исследования. Humanitates. 2018. Т. 4, № 1. С. 43–71.

Ляпунова Ю. И. Концепты основных социальных институтов в Гражданском кодексе Франции и их русские соответствия: Выпускная квалификационная работа. СПб.: БГТУ, 2010. 122 с.

Миллер Дж. А. Магическое число семь плюс-минус два. О некоторых пределах нашей способности перерабатывать информацию. 2010. URL: http://www.ebbinghaus.ru/wpcontent/uploads/2010/02/Miller_564-580.pdf.

Мокиенко В. М., Никитина Т. Г., Николаева Е. К. Большой словарь русских пословиц. М.: ОЛМА Медиа Групп, 2010. 1026 с.

Найшуль В. А. Атлас Букваря городской Руси. β-версия, 2006, 67 с. (рукопись).

Найшуль В. А., Чебанов С. В. Социальная метадисциплина – формальная институционалистика // Третий Российский культурологический конгресс с международным участием «Креативность в пространстве традиции и инновации»: Тез. докл. СПб.: Эйдос, 2010. С. 423–424.

Петров Т. Г. Метод *RHA* как решение проблемы систематизации аналитических данных о вещественном составе геологических объектов // Отечественная геология. 2008. № 4. С. 98–105.

Петров Т. Г., Фарафонова О. И. Информационно-компонентный анализ. Метод RHA. СПб.: Изд-во СПбГУ, 2005. 168 с.

Пущин С. Л. О трех теоремах Б. И. Кудрина // Ценологические исследования. М.: Техника, 2014. Вып. 53. С. 21–28.

Семёнова Д. М. Семантическая структура жестов в первом томе романа Л. Н. Толстого «Война и мир»: Выпускная квалификационная работа. СПб.: БГТУ, 2012. 104 с.

Семёнова Д. М., Чебанов С. В. Ценоз описаний кинесики романа Л. Н. Толстого «Война и мир» // Ценологические исследования. М.: Тех-

ника, 2012. Вып. 46: Специфика ценологических представлений разных школ. С. 181–203.

Смирнова М. А. Концепты социальных институтов в русской сатирической прозе (на примерах произведений М. Е. Салтыкова-Щедрина и В. Н. Войновича): Выпускная квалификационная работа. СПб.: БГТУ, 2008. 70 с.

Фуфаев В. В. Структурно-топологический анализ динамики сообщества банков России в условиях финансового кризиса // Ценологические исследования. Абакан: Центр системных исследований. 2009. Вып. 35: Технетика и ценология: от теории к практике. С. 139–146

Фуфаев В. В. Основы теории динамики структуры техноценозов // Ценологические исследования. Абакан: Центр системных исследований, 1996. Вып. 1: Математическое описание ценозов и закономерности технетики. С. 156–193.

Чебанов С. В. Полнотекстовые базы данных как инструмент понимания (на материале русской лингвосоциологии) // Понимание и рефлексия в коммуникации, культуре и образовании: Материалы Междунар. науч.-практ. Интернет-конференции, посвящ. 70-летию факультета иностранных языков и международной коммуникации Тверского государственного университета. Тверь, 2012. С. 185–197.

Чернышова А. П. Описание концептов социальных институтов в русских народных сказках: Выпускная квалификационная работа. СПб.: БГТУ, 2008, 71 с.

Чеснокова В. Ф. Язык социологии. М.: ОГИ, 2010. 544 с.

Щепаньский Я. Элементарные понятия социологии. М.: Прогресс, 1969. 240 с.

Anderson Ch. The Long Tail: Why the Future of Business Is Selling Less of More. N. Y.: Hyperion, 2006. 238 p.

Babariko M., Jinfeng L., Chebanov S. Idealized Cognitive Model (ICM) of Numbers in the Chinese (C) and Russian (R) Linguistic World Picture (LWP) as a Basis of Conceptual Mapping // 3rd International Congress of Numanities (ICoN 2016). Creativity, Diversity, Development. Program and abstracts. Kaunas, International Semiotics Institute, Kaunas University of Technology, 2016. P. 43–45.

中国谚语资料, 中国文艺研究会资料室主编, 兰州艺术学院文学系 55 级民间文学小组, 上中下三册,上海文艺出版社, 1961 年 1111 页 (Собрание китайских пословиц / Китайская научная библиотека искусств, фольклорная группа факультета литературы Ланьчжоуского института искусств. Шанхай: Изд-во Шанхайской литературы и искусств, 1961. Т. 1, 2. 1111 с.)

Article metadata

Title: Corpus with Specialized Tagging for Studying Statistics of Concepts

Authors: Lin Jinfeng [1], D. M. Semenova [2], S. L. Pushchin [3], T. G. Petrov [4], M. N. Babariko [5], S. V. Chebanov [6]

Authors' e-mail: [1] linjinfeng1990@163.com, [2] dasha.glc@gmail.com, [3] z1q813@mail.ru, [4] tomas_petrov@rambler.ru, [5] maxbabaro@gmail.com, [6] s.chebanov@gmail.com

Authors' affiliation: [1] Lanzhou University of technology (China), [2] LLC "Intelliger" (St. Petersburg, Russian Federation), [3] LLC "Tipografiya KSI-Print" (St. Petersburg, Russian Federation), [4] LLC "Sokolov" (St. Petersburg, Russian Federation), [5] St. Petersburg State University of Economics (St. Petersburg, Russian Federation), [6] St. Petersburg State University (St. Petersburg, Russian Federation)

Abstract. The study of concept statistics involves working with tagging corpuses. In principle, such a tagging can only be manual tagging based on expert assessments involving several experts. However, in some cases this possibility is excluded and the tagging is made by one annotator – the author of the study. The explication of the principles of tagging and reproducible quantitative patterns (covering 80 % of the use of concepts 7 ± 2 of them) suggest that such tagging is satisfactory.

Keywords: concept, concept distribution, quantitative conceptology, manual text tagging, Pareto ratio, Miller magic number.

DOI 10.25205/2307-1737-2020-2-87-113

Reference literature (*in transliteration*):

Anderson Ch. The Long Tail: Why the Future of Business Is Selling Less of More. New York, Hyperion, 2006, 238 p.

Babariko M. N., Chebanov S. V. Russian paremiological arithmology of 19th – 21st centuries. In: Structural and applied linguistics. St. Petersburg, SPbSU Press, 2015, iss. 11, p. 186–219. (in Russ.)

Babariko M. N., Chebanov S. V. The arithmology of Russian Proverbs and sayings of the collection of V. I. Dal. In: Structural and applied linguistics. St. Petersburg, SPbSU Press, 2014, iss. 10, p. 70–91. (in Russ.)

Babariko M., Jinfeng L., Chebanov S. Idealized Cognitive Model (ICM) of Numbers in the Chinese (C) and Russian (R) Linguistic World Picture (LWP) as a Basis of Conceptual Mapping. In: 3rd International Congress of Numanities (ICoN 2016). Creativity, Diversity, Development. Program and abstracts. Kaunas, International Semiotics Institute, Kaunas University of Technology, 2016, p. 43–45.

Chebanov S. V. Full-text databases as a tool of understanding (on the material of Russian linguosociology). In: Understanding and reflection in communication, culture and education: materials of the International scientific and practical Internet-conference dedicated to the 70th anniversary of the Faculty of foreign languages and international communication of Tver state University. Tver, Tver state University, 2012, p. 185–197. (in Russ.)

Chernyshova A. P. Description of the concepts of social institutions in Russian folk tales. St. Petersburg, BSTU Press, 2008, 71 p. (in Russ.)

Chesnokova V. F. Language of sociology. Moscow, OGI, 2010, 544 p. (in Russ.)

Collection of Chinese Proverbs. Chinese scientific library of arts, folk literature group of Lanzhou Institute of arts. Shanghai, Publishing house of Shanghai literature and arts, 1961, vol. 1–2, 1111 p. (in Chin.)

Dal V. I. Proverbs of the Russian people. Moscow, In the University printing house, 1862, 883 p. (in Russ.)

Fufaev V. V. Structural-topological analysis of the dynamics of the community of Russian banks in the financial crisis. In: Coenological research. Abakan, Center for Systems Research, 2009, vol. 35: Technetics and coenology: from theory to practice, p. 139–146. (in Russ.)

Fufayev V. V. Fundamentals of the theory of the dynamics of the structure technocenosis. In: Coenological research. Abakan, Center for system research, 1996, iss. 1: Mathematical description of coenoses and laws of technetics, p. 156–193. (in Russ.)

Karamnov A. S. Quantitative assessment of repetition and complexity of vocabulary in the corpus of the book in English. *Philological sciences. Theory and practice*, 2014, no. 6 (36), part 1, p. 82–86. (in Russ.)

Kasyanova K. About the Russian national character. Moscow, Academic Project; Ekaterinburg, Business Book, 2003, 560 p. (in Russ.)

Kirilova M. V. Conceptualization of social institutions in "Soboryany" N. S. Leskov. Final qualifying work. St. Petersburg, BSTU Press, 2008, 105 p. (in Russ.)

Kirilova M. V., Chernyavskiy V. A. Concepts of social institutions in the work of N. S. Leskova "The cathedral clergy". Institute of sociology RAS. Methodological seminar in memoryof G. S. Batygin, 2009. URL: http://www.isras.ru/files/File/Seminar/Seminar_Batygin/Kirillova_Chernyavsky.pdf (in Russ.)

Kudrin B. I. My seven differences from Zipf. *General and Apply Coenology*, 2007, no. 4, p. 25–33. (in Russ.)

Kurochkina A. S. Social concepts in the linguistic creativity of children in primary school age (diachronic analysis). Final qualifying work. St. Petersburg, BSTU Press, 2008, 241 p. (in Russ.)

Lin Jinfeng, Pushchin S. L., Petrov T. G., Semenova D. M., Chebanov S. V. Truncated Zipf-like distributions and frequency dictionaries of concepts. In: General and applied coenology as an acceptance of the understanding of the fundamental nature of the natural law of the species diversity of individuals in communities in the third scientific picture of the world of material and ideal realities. Practical research. Generalizing materials on general and applied coenology. Proceedings of the XXII meeting-seminar of coenologists (Moscow, NRU MEI, 11.16.2018). Coenological studies. St. Petersburg, KSI-Print, 2019, vol. 59, p. 108–121. (in Russ.)

Lin Jinfeng, Chebanov S. V. Formation of concepts [BODY], [SOUL], [SPIRIT] in the modern Russian language picture of the world. *Bulletin of Tyumen state University. Humanitarian research. Humanitates*, 2018, vol. 4, no. 1, p. 43–71. (in Russ.)

Lin Jinfeng. Concepts [BODY], [SOUL], [SPIRIT] in Russian and Chinese language worldview (anthropological trichotomy in the proverbial worldview). Cand. of Philol. Sci. Diss. St. Petersburg, SPbSU Press, 2018, vol. 1–2, 221 + 147 p. (in Russ.)

Lyapunova Yu. I. Concepts of the main social institutions in the French Civil Code and their Russian compliance. Final qualifying work. St. Petersburg, BSTU Press, 2010, 122 p. (in Russ.)

Mayevskaya A. A. US Constitution Preamble Concepts. Graduation qualifying work. St. Petersburg, BSTU, 2009. (in Russ.)

Miller J. A. Magic number seven plus or minus two. Some limits on our ability to process information, 2010. URL: http://www.ebbinghaus.ru/wpcontent/uploads/2010/02/Miller_564-580.pdf. (in Russ.)

Mokienko V. M., Nikitina T. G., Nikolaeva E. K. Large dictionary of Russian Proverbs. Moscow, OLMA Media Group, 2010, 1026 p. (in Russ.)

Naishul V. A. Atlas of the Primer of urban Russia. β-version, 2006. 67 p. (manuscipt). (in Russ.)

Naishul V. A., Chebanov S. V. Social metadiscipline – a formal institutionalistic. In: The 3rd Russian cultural congress with international participation "Creativity in the space of tradition and innovation". Thesis of reports. St. Petersburg, Eidos, 2010, p. 423–424. (in Russ.)

Petrov T. G., Farafonova O. I. Information-component analysis. RHA-method. St. Petersburg, SPbSU Press, 2005, 168 p. (in Russ.)

Pushchin S. L. On three theorems of B. I. Kudrin. In: Coenological research. Moscow, Technique, 2014, iss. 53, p. 21–28. (in Russ.)

Semenova D. M., Chebanov S. V. Coenosis of descriptions of kinesics of the novel L. N. Tolstoy's "War and peace". In: Coenological research. Moscow, Technique, 2012, iss. 46: Specific of coenological views of different schools, p. 181–203. (in Russ.)

Semenova D. M. Semantic structure of gestures in the first volume of the novel by L. N. Tolstoy's "War and peace". St. Petersburg, BSTU Press, 2012, 104 p. (in Russ.)

Shchepanskiy J. Basic concepts of sociology. Moscow, Progress, 1969, 240 p. (in Russ.)

Smirnova M. A. Concepts of social institutions in Russian satirical prose (on the examples of works by M. E. Saltykov-Shchedrin and V. N. Voinovich). Final qualifying work. St. Petersburg, BSTU Press, 2008, 70 p. (in Russ.)

Zakharov V. P., Bogdanova S. Yu. Corpus linguistics: Book for students of the direction "Linguistics". St. Petersburg, SPbSU Press, 2013, 148 p. (in Russ.)