ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 252 за 2018 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Кислицын А.А., Орлов Ю.Н.

Структура семантических связей слов русского толкового словаря

Рекомендуемая форма библиографической ссылки: Кислицын А.А., Орлов Ю.Н. Структура семантических связей слов русского толкового словаря // Препринты ИПМ им. М.В.Келдыша. 2018. № 252. 28 с. doi:10.20948/prepr-2018-252
URL: http://library.keldysh.ru/preprint.asp?id=2018-252

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.А. Кислицын, Ю.Н. Орлов

Структура семантических связей слов русского толкового словаря

Москва — 2018

Кислицын А.А., Орлов Ю.Н.

Структура семантических связей слов русского толкового словаря

В работе представлен результат статистического анализа графа связей между словами в соответствии со статьями толкового словаря русского языка. Выяснено, что граф словаря содержит сильно связную компоненту (ядро), через которую в конечном счете осуществляется объяснение смысла слов, самоподобную по типам связей оболочку ядра (канал) и периферию, в которой наряду со связями с ядром присутствуют в малой доле внутренние циклы. Исследована структура связей ядра и выделены основные кластеры, с помощью которых предлагается проводить семантический анализ текста.

Ключевые слова: толковый словарь, сильно связная компонента, кластеризация, семантический анализ

Kislitsyn A.A., Orlov Yu.N.

Structure of semantic connections of words of the Russian explanatory dictionary

In this paper the result of statistical analysis of connections between words of the Russian explanatory dictionary is presented. It is found that the dictionary graph contains a strongly connected component, through which the explanation of the meaning of words is ultimately carried out. This graph also contains a self-similar shell of the kernel, so-called channel, and the periphery in which, along with the kernel connections, internal cycles are exist in a small fraction. The structure of the core links is studied and the main clusters are identified, with the help of which it is proposed to carry out semantic analysis of the text.

Key words: explanatory dictionary, strongly connected component, clusterization, semantic analysis

Работа выполнена при поддержке гранта РФФИ № 18-01-00619.

Содержание

Введение и постановка задачи 3
1. Общая структура связей слов в словаре 5
2. Структура ядра 8
3. Основные тематические кластеры ядра 14
4. Оценка близости смыслов высказываний 26
Литература 28

Введение и постановка задачи

Настоящая работа продолжает исследование статистических закономерностей онтологических связей ориентированных графов с большим количеством вершин и сильно разреженной матрицей смежности, предпринятое в [1, 2]. В качестве анализируемого объекта рассматривается толковый словарь русского языка [3]. Применительно к словарю граф связей строится в соответствии со словарными статьями, в которых объясняется значение того или иного слова.

Целью работы является изучение структуры связей слов в словаре с последующим использованием этой структуры для формализации семантического анализа текста, т.е. для машинного обучения при решении задачи определения смысла слова или высказывания.

Процесс объяснения смысла слова с помощью статей толкового словаря состоит в следующем. Стартуя с какого-либо слова (нулевой уровень иерархии объяснения применительно к данному слову), мы читаем объясняющую его словарную статью (первый уровень иерархии), после чего каждому слову в этой статье ищем объяснение в том же словаре (второй уровень иерархии) и т.д. Предполагается, что в словаре не используются слова, значения которых не объяснены. Процедура объяснения заканчивается, когда каждая ветвь заворачивается в цикл, т.е. когда на определенном уровне иерархии не возникает слов, на которые не было ссылок с предыдущих уровней.

Возникают следующие вопросы.

1. Все ли слова словаря взаимно связаны между собой словарными статьями, или же существует один или несколько базисных кластеров, из которых не выходит ни одной связи?

2. Если в словаре существует сильно связный подграф, то какова статистика связей внутри вершин этого подграфа, а также связей остальной части словаря со своей сильно связной компонентой?

3. Какова глубина иерархии словаря по замыканию набора объясняющих слов?

4. Каково распределение циклов сильно связной компоненты по длинам?

Для удобства анализа была сделана следующая предобработка словаря. Из него были исключены все союзные слова, а прилагательные, наречия и глаголы заменены образующими их существительными без указания смысловой иерархии внутри одной словарной статьи. Например, слово «местность» определяется в словаре как «1. Какое-нибудь определенное место, пространство, участок земной поверхности» с примерами использования этого слова: «гористая местность, открытая местность, занятия на местности», а также как «2. Край, округа» с примерами: «дачная местность, в нашей местности». При построении онтологии все поясняющие примеры были исключены. Объясняющие слова были объединены в одну иерархию первого уровня, так что получилось:

«*местность*» = «*определенность, место, пространство, участок, земля, поверхность, край, округа*».

Как видно, в одном уровне иерархии могут оказываться как синонимы, используемые для объяснения данного слова, так и некоторые общезначимые слова-связки: пространство, земля и т.п. Далее каждое слово иерархии первого уровня представлялось в аналогичном виде, так что по отношению к слову «местность» (в данном примере это есть порождающее слово или слово нулевого иерархического уровня) выстраивалась иерархия второго уровня.

Заметим, что если в словарной статье встречалось само объясняемое слово, оно исключалось из нее. Это сделано потому, что использование в определении (первый уровень) самого определяемого слова (нулевой уровень) является логической ошибкой. Можно, конечно, считать, что смысл подобных слов в данном словаре не определен, но такая позиция неконструктивна. Например, в определении «ткань» = «тканая материя» слово «тканая» исключается, так что остается только «*ткань*» = «*материя*». Таким образом, из фактической онтологии словаря были исключены циклы нулевой длины.

Также отметим, что на данном уровне иерархии каждое слово указывается только один раз безотносительно к тому, сколько раз оно использовалось в рассматриваемой словарной статье. Например, «колода» = «1. Короткое толстое бревно», «2. Предмет, представляющий собою бревно», «3. Комплект игральных карт». Слово «бревно» участвует в объяснении дважды, но учитывается в схеме только один раз. В итоге получается, что «*колода*» = «*краткость, толстота, бревно, предмет, представление, комплект, игра, карта*». Вообще, каждое слово упоминается в онтологии только один раз. Частота его использования отражается числом ориентированных ребер, входящих в соответствующую вершину.

В соответствии с описанной структурой словарных статей смыслом слова следует считать не собственно словарную статью, а совокупность слов (статей), связанную ориентированными ребрами так, что в этой совокупности существует последнее слово, т.е. такое слово, которое не связано с новыми вершинами. Следовательно, таким последним словом может быть только само первоначально объясняемое слово. В результате получаем, что в терминах онтологического графа **смысл слова следует трактовать как упорядоченную совокупность слов словаря, образующую циклический подграф в направленном графе связей.**

Проиллюстрируем это определение на примере самого слова «смысл». В соответствии со словарем [3], слово «смысл» определяется следующим образом: «Смысл: содержание, сущность, суть, значение чего-нибудь; цель, разумное основание чего-нибудь; разум, разумность». Заметим, что это слово входит в ядро словаря (см. далее п.1), внутри ядра оно имеет 20 входящих связей. Согласно вышеприведенному словарному определению слова «смысл» можно установить связь «смысл – значение». Переходя, в свою очередь, к слову «значение», читаем: «Значение: смысл, то, что данное понятие, явление, предмет значит, обозначает». Таким образом, существует и обратная связь

между этими словами, т.е. «значение – смысл». В результате имеем цикл минимально возможной длины 2 (по количеству ребер получившегося подграфа) «смысл – значение – смысл». Подчеркнем, что циклы длины 1 невозможны по определению цикла, а циклы длины 0 (петли) исключены из рассмотрения, поскольку слова не могут объясняться через самих себя. Кроме приведенной, существуют и другие циклические связи слова «смысл». В частности, поскольку «значение» связано со словом «понятие», которое, в свою очередь, связано с другими словами соответствующими статьями словаря, то можно обнаружить, например, и такой цикл: «смысл – значение – понятие – мысль – идея – замысел – смысл». Длина этого цикла равна шести. Аналогично поступаем и для других слов, входящих в первый уровень объяснения.

Совокупность циклов, связанных с данным словом, образует его смысловую категорию. При этом часть циклов может относиться к одному понятийному кластеру, а другая часть – к другому, поскольку слова могут иметь несколько значений. Следовательно, ключевой задачей является построение распределения слов словаря по связанным с ними циклам. Частично эта задача была исследована в [1]. В настоящей работе приводятся результаты, уточняющие полученные ранее.

1. Общая структура связей слов в словаре

В работе [1] было выяснено, что в толковом словаре [3] существует **ядро,** т.е. **сильно связная компонента**, в которой для любых двух ее вершин можно найти ориентированный путь через некоторые другие вершины, связывающий их между собой. Ядро словаря строилось в [1] для слова «предмет», взятого в качестве нулевого уровня объясняющей иерархии.

На рис. 1 приведен пример построения направленных связей для иерархии, связанной со словом «предмет», чтобы проиллюстрировать процесс нахождения матрицы смежности для ядра словаря. Первый уровень иерархии слова «предмет» содержит 10 слов, все они находятся во второй колонке рис. 1. Затем для некоторых слов первого уровня, выделенных жирным шрифтом, строятся аналогичные связи-стрелки. Появляющиеся при этом новые слова, не использовавшиеся ранее, помещаются в следующий иерархический столбец.

Ядро словаря оказалось состоящим из 3874 слов, распределенных по 19 уровням иерархии, включая нулевой, состоящий из одного слова. Таким образом, объясняющих иерархий, в которых появляются новые слова, всего 17. В последнем уровне иерархии ядра уже не содержится новых слов по сравнению с предыдущими уровнями, так что процесс объяснения на этом заканчивается. При этом в уровнях иерархии с 11-го по 17-ый содержится всего 74 слова (это менее 2% слов ядра), появление которых относительно случайно и зависит от выбора синонимов в объясняющих словах предыдущих иерархий. Это есть определенная погрешность в формировании ядра, к обсуждению которой мы вернемся в п. 2.

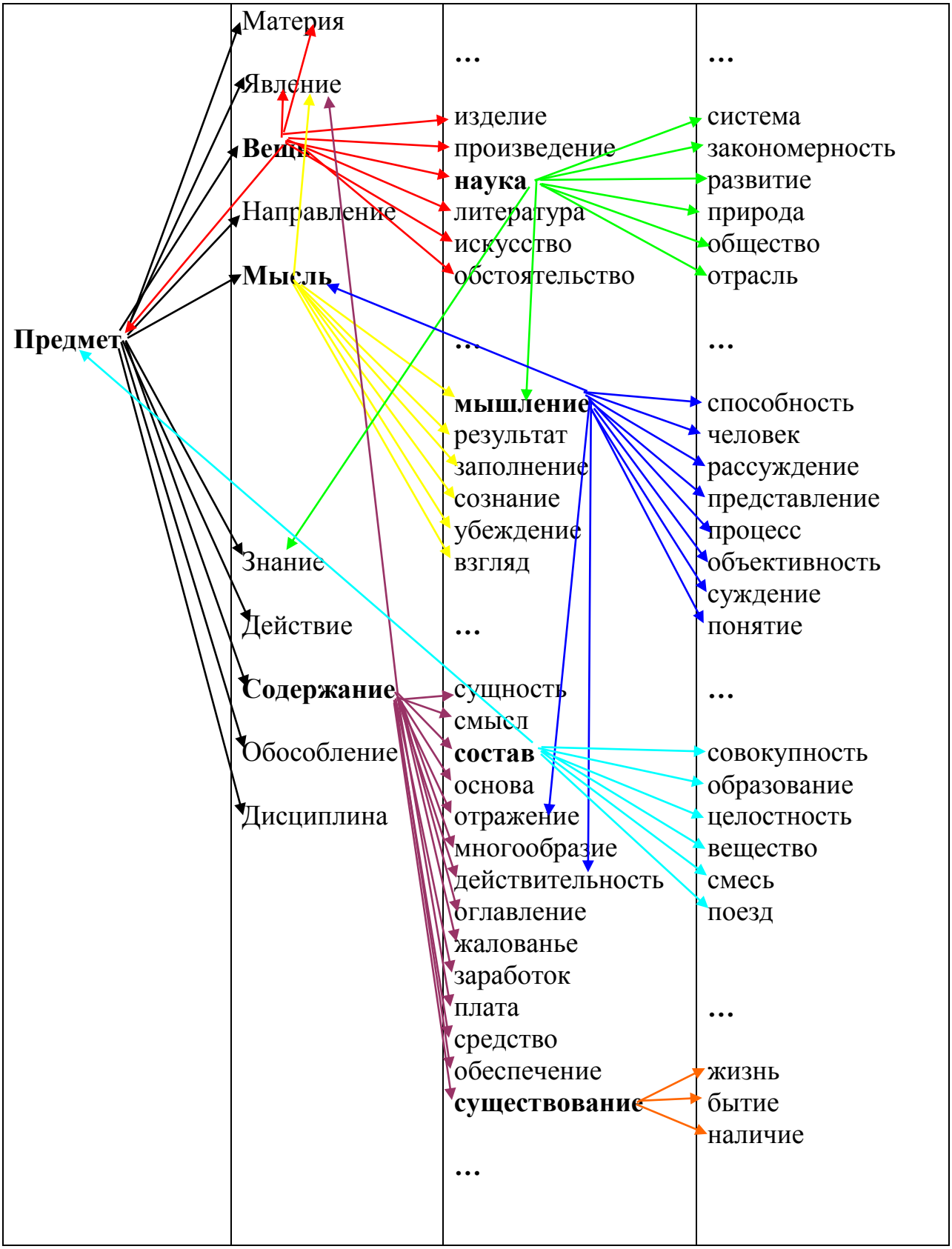

Рис. 1. Пример построения иерархии для слова «предмет»

В словаре [3] содержится 80 тыс. слов, т.е. словарных статей, из которых около половины (35 тыс. слов) относится к существительным, для которых и строится граф онтологических связей. Таким образом, ядро составляет по объему примерно 10% словаря в части существительных. Подчеркнем, что из ядра не выходит вовне ни одной связи. Остальные 90% слов в идеальном словаре должны объясняться через ядро. Идеальность предполагает отсутствие циклов вне ядра. Указанные циклы возникают тогда, когда составитель недостаточно четко использует синонимы для объяснения смысла слов. Поскольку выделение ядра и построение графа связей – весьма трудоемкая статистическая процедура, она, естественно, не может быть проведена только методами лингвистики. Поэтому реальные словари не идеальны, и вне ядра также содержится некоторое количество циклов. Применительно к словарю [3] таких слов, имеющих циклические связи вне ядра, оказалось 1925, т.е. приблизительно 5% словаря. Эти слова будем называть **периферией.** Оставшиеся 85% слов образуют определенную структуру связей с ядром словаря (рис. 2), называемую **каналом.** Канал образован **оболочкой ядра**, в которой содержатся слова, связанные только с ядром (выделены жирным на рис. 2), и **каскадами** разных уровней, слова в которых связаны только с каскадами низших уровней, оболочкой и ядром.

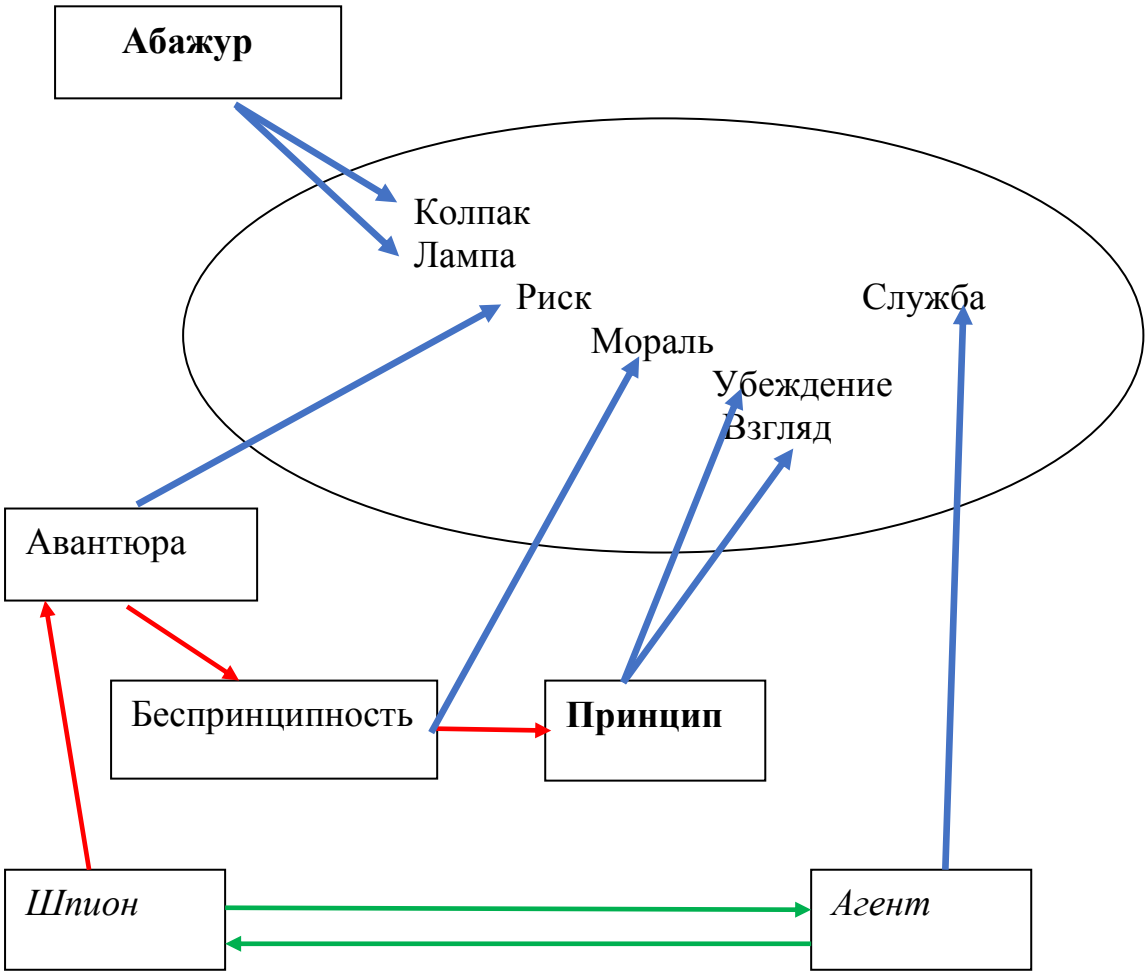

Рис. 2. Фрагмент графа связей, порожденных словами извне ядра

На рис. 2 ядро обозначено овалом, с которым связаны все остальные слова. Эти связи показаны синими стрелками. Слова из оболочки ядра (в примере на рис. 2 это «абажур» и «принцип») связаны только с ядром. Существует также каскад (первого уровня), слова из которого связаны только с оболочкой и ядром. Это, например, слово «беспринципность». Существует каскад второго уровня, слова из которого (например, «авантюра») связаны с каскадом первого уровня, оболочкой и ядром. Связи в каскадах показаны красными стрелками. Периферия же характеризуется тем, что ее слова связаны не только с каналом и ядром, но и между собой (курсив и зеленые стрелки).

Интересно отметить определенное **самоподобие** структуры канала. Именно, отношение количества слов, содержащихся в оболочке и каскадах разных уровней, связано между собой степенью двойки. Из примерно 29 тыс. слов вне ядра и периферии примерно половина (15 тыс.) находится в оболочке. Каскад первого уровня содержит 7,3 тыс. слов (т.е. примерно половину остатка), в каскаде второго уровня находится 3,5 тыс. слов (половина оставшейся половины) и т.д. Всего канал содержит 8 каскадов, в последнем находится 46 слов. Какие именно слова содержатся в последнем каскаде, не имеет особенного значения, поскольку, как уже говорилось выше, периферия в значительной степени случайна и при незначительной модификации определяющих статей может перейти в тот или иной каскад.

Удобство выделения каскадов состоит в том, что связи в них имеют только одно направление – в сторону ядра; параллельных связей (внутри данного уровня) и восходящих к высшим уровням нет. Следовательно, для каждого слова можно априори установить область словаря, в которую из этого слова не выходит ни одного связующего ребра.

2. Структура ядра

Рассмотрим теперь структуру связей ядра. Предварительно следует сделать несколько уточнений.

Во-первых, в силу уже упоминавшейся нечеткости выбора синонимов в ядре словаря существуют **тупиковые циклы**. Цикл считается тупиковым, если хотя бы одна из вершин цикла связана только с одной вершиной того же цикла. Предположим, что на каком-то этапе возникает слово «X», которое объясняется словами «A» и «B», а затем обнаруживается, что «A» объясняется только через «B» или через «B» и «X». Характерным примером такого рода является слово «липкость», определяемое как «прилипание, клейкость», а «клейкость», в свою очередь, определяется как «липкость». Однако доля таких циклов в ядре очень мала (менее 1 % слов), поэтому можно вручную исправить соответствующее определение. Например, если определить «клейкость» как «способность вещества или тела приклеиваться или прилипать к поверхности», то вершинами графа для слова «клейкость» будут «способность, вещество, тело, приклеивание, прилипание, поверхность», и тупиковый цикл разрушится.

Во-вторых, в словаре содержатся не вполне корректные определения некоторых понятий, имеющих наряду с общим и специальное значение. В результате формируется искаженный граф связей между словами. То, какие именно слова и в каком количестве войдут в ядро словаря, зависит в том числе и от точности даваемых определений. Например, статья «квадратура» дается в контексте выражения «квадратура круга» и обозначает следующее: «неразрешимая задача превращения круга в равновеликий квадрат». Это, строго говоря, неверно, поскольку данная задача не вообще «неразрешима», а не имеет решения с применением только циркуля и линейки. Используемое слово «превращение» еще более запутывает и не слишком подходит для объяснения математического термина, причем в рамках самого словаря оно означает «придание иного вида», что тем более не позволяет понять проблему неразрешимости: круг ведь можно растянуть и деформировать до состояния квадрата. Другой пример – словарная статья для слова «предел». Она имеет три подраздела: два общих (страна, местность; крайняя степень чего-либо в контексте «дойти до предела возможностей») и один математический. Последний звучит так: «пространственная или временная граница чего-либо». То, что это самое «что-либо» должно присутствовать в любой ненулевой окрестности упомянутой «границы», из данного определения не следует, что искажает строгую математическую формулировку. Тем не менее, такие словарные статьи не корректировались. Все это приводит к тому, что ядро известно с точностью порядка 5 %, если учесть также возможность попадания в него слов с периферии после расцепления периферических циклов.

Рассмотрим сначала статистику вершин и связей в зависимости от уровня иерархии в объяснении исходного слова с нулевого уровня (рис. 3).

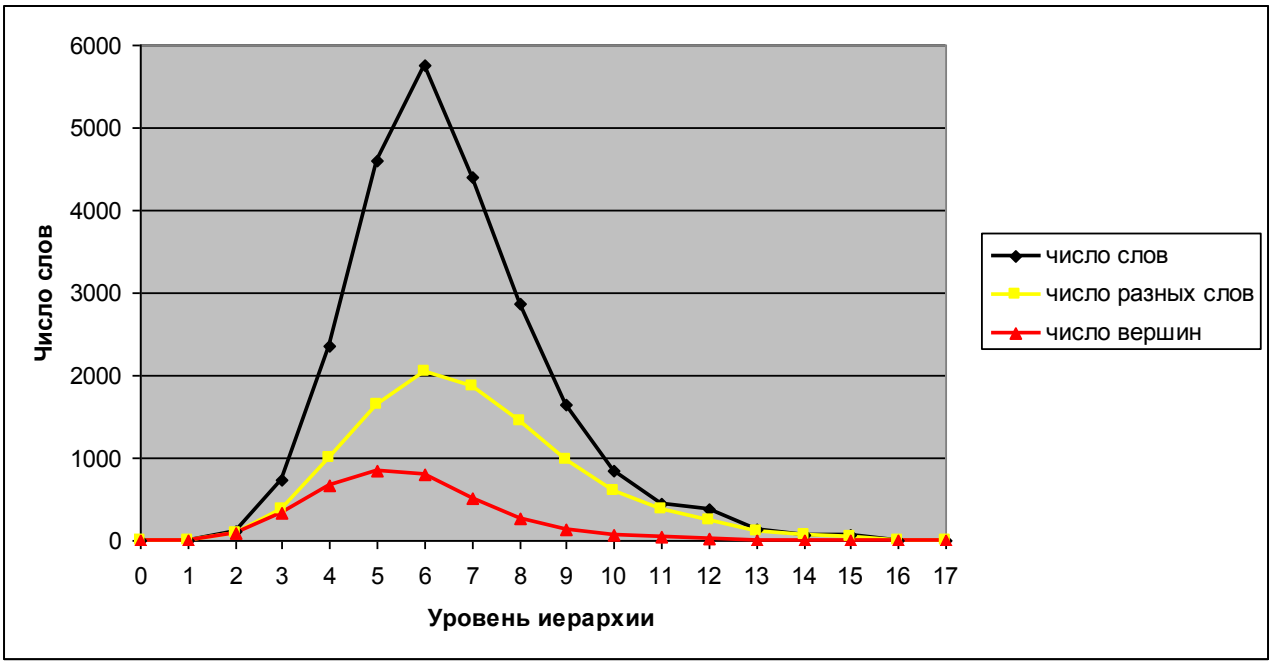

Рис. 3. Распределение числа слов по уровням иерархии

В легенде рис. 3 «число слов» – это полное число входящих ребер в иерархию данного уровня, т.е. это число слов, участвующих в объяснении слов с предыдущего уровня; «число разных слов» – это число вершин, с которыми соединяются выходящие с предыдущего уровня ребра, а «число вершин» – это число новых слов, появившихся на данном уровне и отсутствовавших на предыдущих уровнях, т.е. слов, не использовавшихся ранее.

Применительно к слову «предмет» на первом уровне иерархии все три показателя числа слов на рис. 3 совпадают и равны 10, т.е. в объяснении слова «предмет» участвуют 10 новых слов. На следующем уровне иерархии, порожденном объяснением этих 10 слов, используются уже 88 слов, из которых новыми, т.е. не использовавшимися ранее, являются 84 слова, и ко всем этим 88 словам приходят 107 связей с первого уровня. Аналогичный смысл имеют и остальные точки графиков на рис. 3.

Наибольшее число слов, а также и наибольшее число ребер появляется на шестом уровне объяснения смысла данного слова, порождающего ядро. Число вершин, т.е. число новых слов, появляющихся на определенном уровне иерархии, имеет максимум несколько раньше – на пятом уровне. Заметим также, что число вершин приблизительно совпадает с числом ребер для первых трех уровней иерархии, после чего число ребер (т.е. полное число слов, черная кривая на рис. 3) резко возрастает. Это означает, что количество новых смыслов на четвертом уровне иерархии по отношению к данному слову нулевого уровня становится очень велико, так что вряд ли целесообразно удерживать их всех в памяти (реального человека или компьютера) вне связи с контекстом. Далее в п. 3 будет рассмотрен способ кластеризации сильно связной компоненты на условно несвязные кластеры с ограничением на длину циклов. На рис. 4 приведено распределение степеней вершин, т.е. вероятностей того, что с некоторым словом ядра связано заданное число других слов ядра.

Рис. 4. Распределение степеней вершин ядра

Совокупная степень 3874 вершин графа ядра по входящим (и, соответственно, выходящим) ребрам составила 25916, т.е. это полное число ребер сильно связной компоненты.

Распределение слов ядра только по числу выходящих ребер представлено на рис. 5. Оно показывает, какова вероятность того, что в объяснении данного слова на следующем уровне иерархии участвует заданное количество слов.

Рис. 5. Распределение числа ребер, выходящих из вершин графа

Наиболее вероятное значение этого распределения равно 4. Среднее число связей с объясняющими словами равно приблизительно 7. Максимальное количество связей 38, минимальное 1.

Наибольшее число связей (по 38) имеют два слова: «ключ» и «форма». Далее по убыванию следуют: 35 «время», «класс», «место», «расположение»; 32 «род»; 31 «петля», «план», «точка»; 30 «партия», «фигура», «фокус»; 29 «спутник»; 28 «мир», «положение»; 27 «язык»; 26 «дело», «отношение», «подведение», «принятие», «святой», «сила»; 25 «дух», «команда», «корень», «крыло», «проведение», «солнце»; 24 «венец», «перспектива», «политика», «работа», «соединение», «стопа»; 23 «голова», «земля», «клуб», «колодка», «лист», «основание», «открытие», «подача», «полоса», «порода», «ракета», «сбор», «система», «тень», «ядро»; 22 «масса», «поле», «практика», «свет», «спуск», «ток», «трубка»; 21 «вал», «ворота», «вывод», «жизнь», «корпус», «линия», «оборот», «образ», «общество», «ось», «положительность», «пункт», «связь», «тяжесть»; «экономика».

Слов, для объяснения которых на следующем уровне иерархии требуется более 20 связей, в ядре всего 72, что составляет менее 2% слов ядра. При этом в интервале от 1 до 12 связей содержится 90% распределения.

На рис. 6 приведено распределение слов ядра по входящим ребрам. Оно показывает, как часто для объяснения слов ядра требуется данное количество связей.

Рис. 6. Распределение слов ядра по числу входящих ребер

Из построенных приблизительно 26 тыс. связей между словами наибольшую частоту использования имеет слово «часть»: оно встретилось 265 раз. Далее по нисходящей идут слова: «человек» (208 раз), «предмет» (180), «отсутствие» (165), «место» (159), «действие» (149), «делание» (131), «движение» (127), «состояние» (113), «жизнь» (103), «вид» (103), «форма» (102), «животное» (100). Это наиболее популярные слова ядра. Их 13, число связей с которыми не менее 100.

Наиболее часто встречается ситуация, когда в данное слово входит всего одно ребро. Таких слов в ядре оказалось 1406, что составило 0,35 от общего числа слов ядра. Эти слова в совокупности будем называть **границей ядра.** Вероятность того, что в некоторое слово входит не более двух ребер, равна приблизительно 0,5, не более трех – соответственно 0,6, не более пяти ребер 0,75, не более 10 ребер 0,85, не более 15 ребер 0,9.

На рис. 7 показано совместное распределение вершин ядра по числу входящих (IN) и выходящих (OUT) ребер. Эти распределения по отдельности

были приведены выше на рис. 5 и 6. Они оказались практически независимыми, между числом входящих и выходящих ребер имеется незначительная положительная корреляция, равная 0,27. Вероятность максимума (1 входящее ребро и 4 выходящих) равна 0,055, что приблизительно равно произведению соответствующих независимых распределений в точках их максимумов.

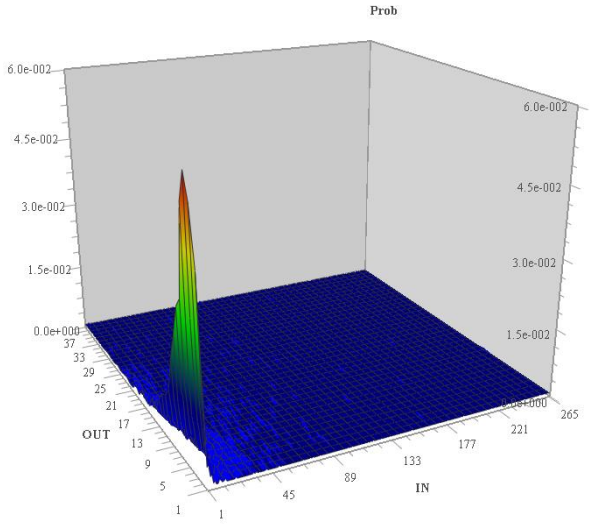

Рис. 7. Совместное распределение вершин по входящим и выходящим ребрам

Выяснилось также, что вероятность связи слова канала с ядром не зависит от номера каскада. Ядро в этом смысле универсально, примерно 90% связей слов каскада любого уровня связано с ядром напрямую. Остальные 10% равномерно распределяются по совокупному объему каскадов низших уровней и оболочки ядра. Важно также отметить, что распределение связей внешних и внутренних слов ядра одинаково: в границу ядра попадает 35% тех связей, которые направлены к ядру из оболочки и каскадов, но эти связи количественно малы (1-2 ребра). Как и для ядра, наиболее востребованные по входящим ребрам слова оказываются столь же часто используемыми и для слов канала: это те же, что и отмеченные ранее, наиболее популярные слова «часть», «предмет», «человек» и т.д.

На рис. 8 показана доля новых слов (т.е. вершин графа), появляющихся на определенном уровне иерархии, по отношению к полному числу выходящих с этого уровня ребер. С высокой точностью указанная зависимость является степенной с отрицательным показателем степени:

$$K_{new}(t) \approx at^{\gamma}, \quad \gamma \approx -3/2.$$

Это означает, что число связей растет с большей скоростью, чем число вершин.

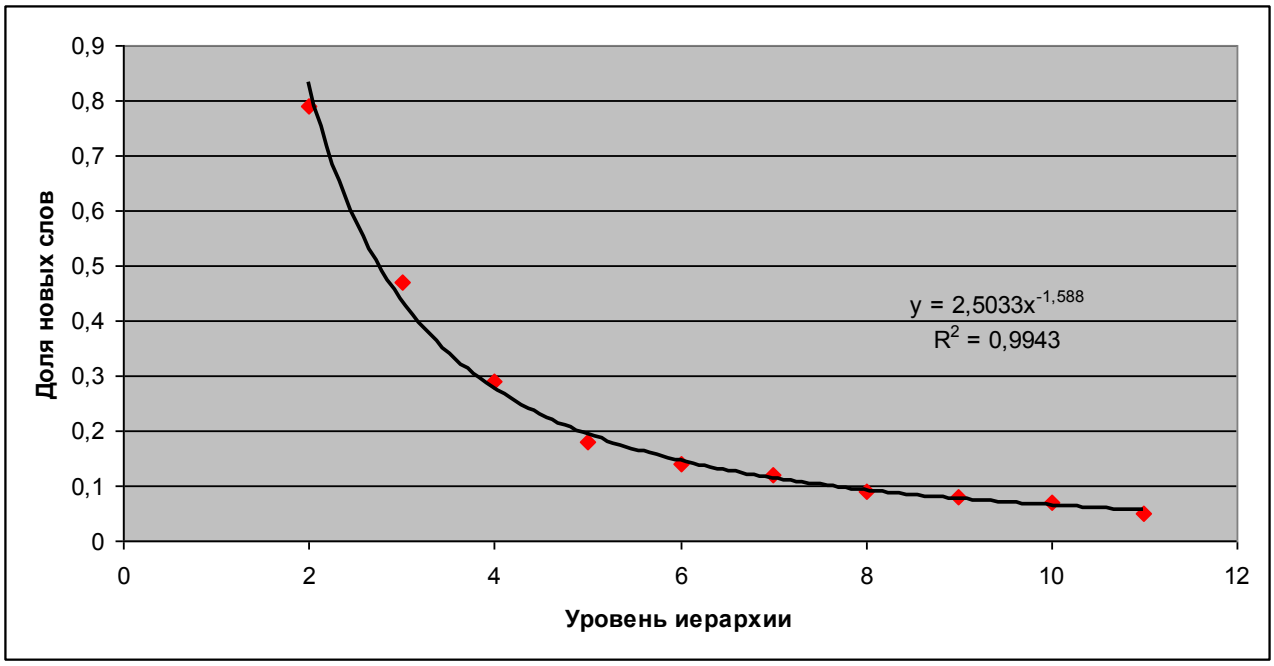

Рис. 8. Доля новых слов в зависимости от уровня иерархии

Отмеченный рост числа связей по сравнению с числом вершин позволяет предположить, что когда рассогласование между этими двумя параметрами становится слишком большим, то новые вершины перестают в полном объеме учитываться в возможных объяснениях значений слов.

3. Основные тематические кластеры ядра

Если занумеровать слова в словаре, например, в алфавитном порядке, то структурно схема ядра может быть представлена матрицей из нулей и единиц (0_{jk} – вершины j, k не соединены, 1_{jk} – имеется направленная связь от вершины j к вершине k) размером $N \times N$, где N есть число слов в ядре. Структурная матрица ядра весьма разрежена, поскольку для примерно 4 тыс. строк (столбцов) в каждой строке этой длины находится в среднем 6 единиц, а остальные элементы – нули.

Для целей нашего исследования необходимо установить, сколько циклов и какой длины существует в сильно связной компоненте. Поскольку петли в ядре отсутствуют, то число циклов длины n равно следу n-ой степени матрицы смежности. Интерес представляют циклы без самопересечений.

Количество циклов в зависимости от их длины представлено в табл. 1. Также в ней указано количество слов ядра нарастающим итогом, которые оказываются охваченными циклами длины 2, 2 и 3, 2, 3, и 4 и т.д. Выясняется, что для полного циклического охвата ядра требуется учет циклов всех длин от 2 до 13. Отметим, что примерно столько же иерархий требуется для взаимного объяснения всех слов ядра. Практический же расчет числа циклов с длиной больше 10 весьма труден, так как требует очень большого перебора данных.

Табл. 1. Зависимость количества циклов от длины цикла

Длина цикла	2	3	4	5	6	7	8
Количество циклов	1014	865	2712	15997	116960	910296	7251518
Количество слов ядра, нарастающим итогом	1416	1808	2181	2636	3091	3452	3646

Количество циклов с увеличением их длин возрастает экспоненциально. Аппроксимационная формула на уровне детерминации $R^2 = 0{,}99$ имеет вид: $n(l) \approx 2{,}16\exp(1{,}85l)$. Максимальное число циклов отвечает длине цикла, равной примерно половине количества вершин графа, оно приблизительно оценивается величиной 10^{1000}. Столько потенциальных смыслов содержится в словах ядра. Для сравнения укажем, что число атомов во Вселенной оценивается всего лишь как 10^{80}.

В короткие циклы длин 2, 3 и 4 вовлечено совокупно около половины слов ядра. Это относится не только к ядру в целом, но и к его границе, которая также представлена в коротких циклах половиной своего состава. На рис. 9 показана структура ядра, если в нем ограничиться рассмотрением только тех слов, которые образуют циклы длины 2.

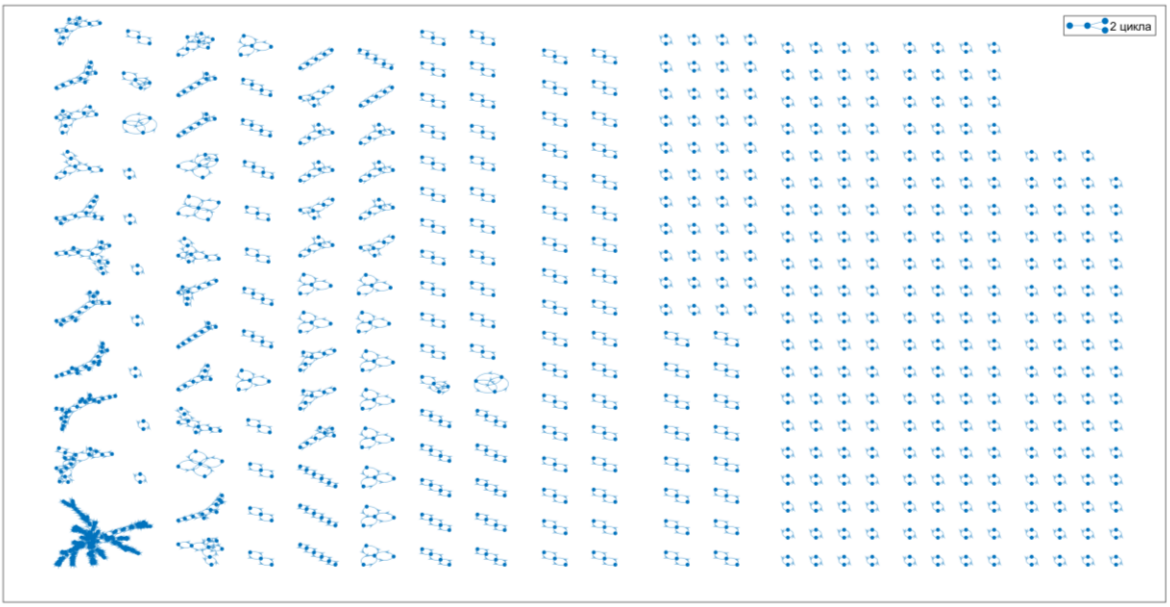

Рис. 9. Структура ядра в циклах длины 2

На рис. 9 видны изолированные циклы длины 2 (справа), далее описаны циклы длины 2, связанные в двухзвенную цепь, трехзвенную и т.д. Заметим, что цепочка циклов имеет естественную интерпретацию как модель ассоциативного мышления, когда смысл, описываемый первым циклом,

приобретает дополнительные черты через общую вершину, связанную со вторым циклом.

В нижнем левом углу рис. 9 показан остаток сильно связной компоненты, состоящей из 198 вершин, образующих 418 циклов длины 2. Этот фрагмент детально показан на рис. 10.

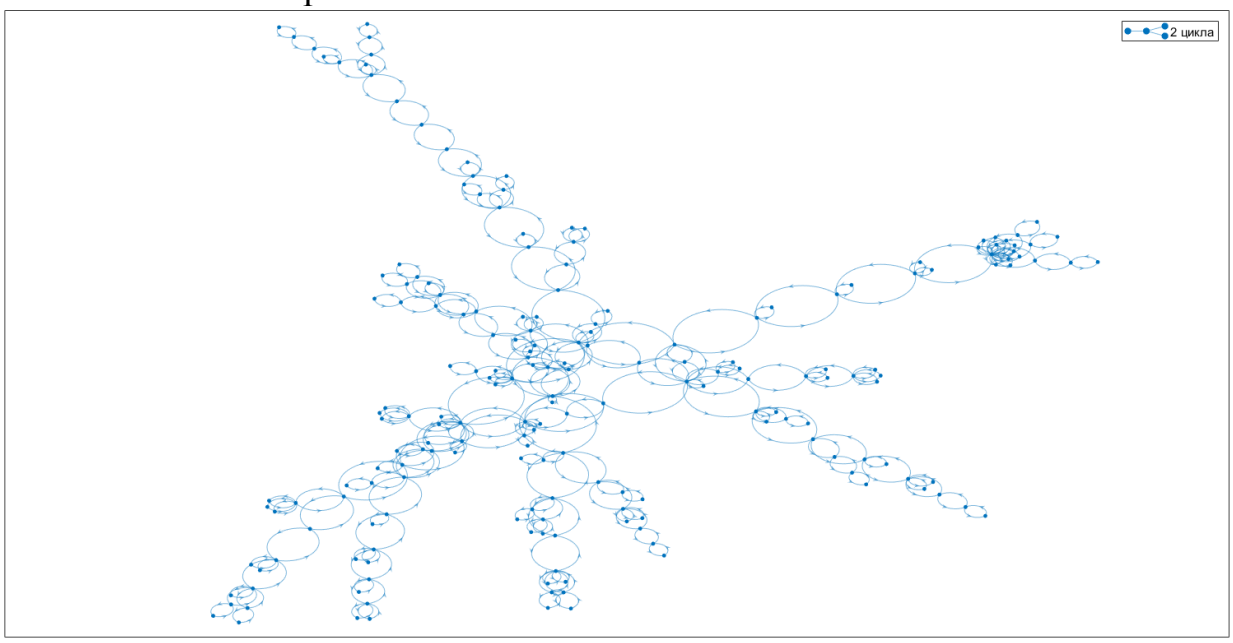

Рис. 10. Остаток сильно связной компоненты в циклах длины 2

Если добавить циклы длины 3, то ядро представится в виде (рис. 11):

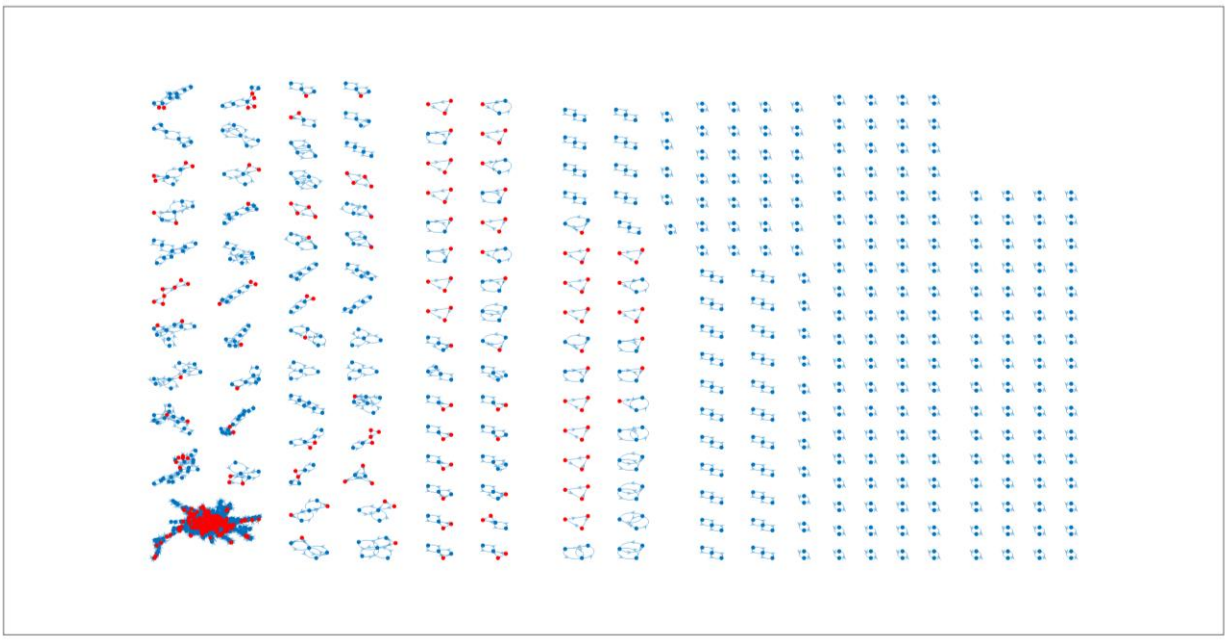

Рис. 11. Структура ядра в циклах длин 2 и 3

Красным цветом на рис. 11 обозначены циклы длины 3.

В целом прослеживается тенденция увеличения числа циклов для тех слов, которые чаще употребляются в ядре, однако явной зависимости числа циклов от числа входящих связей нет. Наибольшее число разных циклов длины 2

имеется у слова «время» (15 циклов), которое имеет 76 входящих ребер. Следующим идет слово «положение» (10 циклов длины 2, 82 входящих ребра), 8 циклов у слова «работа», по 7 циклов имеют слова «общество», «место», «человек». Наибольшее число циклов длины 3 имеет слово «жизнь» при числе входящих ребер 103. В циклах длины 4 чаще всего участвует слово «положение» (265 циклов). В то же время весьма часто используемое слово «форма» (102 входящих ребра) вообще не имеет циклов длин 2, 3 и 4.

В целом же сильно связная компонента словаря имеет вид, показанный на рис. 12. Отчетливо просматривается лишь граница – вершины, имеющие ровно одно входящее и одно выходящее ребро. Остальные связи сильно запутаны.

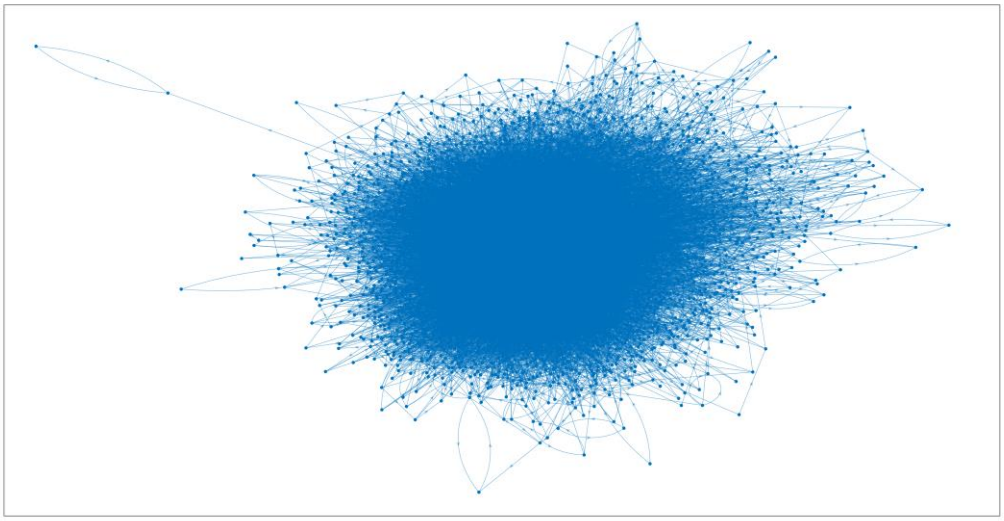

Рис. 12. Сильно связная компонента словаря

Практический интерес представляют тематические кластеры ядра, на которые распадается сильно связная компонента, если ограничиться только циклическими связями длин 2, 3 и 4. Некоторые из них перечислены ниже, в скобках указаны слова, с которыми данное слово связано в ядре (но необязательно в указанном кластере).

Война (55 слов): *командующий* (начальник, крупность, войско, соединение), *копье* (укол, метание, оружие, древко), *мишень* (предмет, служение, цель, учеба, тренировка, стрельба), *плен* (состояние, нахождение, человек, захват, противник, лишение, свобода), *подвиг* (герой, поступок), *пушка* (длина, ствол, артиллерия, орудие, стрельба, дальность, расстояние), *револьвер* (множество, заряд, рука, огонь, стрельба, оружие, магазин, вид, вращение, барабан), *стрелок* (человек, умение, стрельба, военнослужащий, войско, веде'ние, огонь, танк, самолет, охрана), *стычка* (малость, бой, ссора, столкновение), *фланг* (сторона, шеренга, фронт, расположение, войско), *штурм* (приступ, решительность, атака, укрепление, позиция, крепость), *штык* (укол, оружие, прикрепление, конец, ствол, винтовка, ружье), *бомба* (снаряд, бросок, самолет, взрыв, устройство, округлость, полет), *винтовка* (рука, огонь,

стрельба, оружие, винт, нарезка, канал, ствол, надевание, штык), *гильза* (металл, трубка, пуля, снаряд, дно, отверстие, капсюль), *дивизия* (основа, тактика, соединение, различие, вид, вооружение, сила), *неприятель* (враг, войско, противник, человек, враждебность, настроение), *патрон* (пуля, заряд, капсюль, заключение, гильза, полость, труба, цилиндр, приспособление, вставка, заготовка, инструмент, деталь), *приступ* (атака, штурм, острота, внезапность, проявление, признак, болезнь, припадок), *формирование* (воинство, часть, придание, определенность, форма, законченность, порождение, создание, организация, состав), *враг* (человек, нахождение, состояние, вражда, противник, война, неприятель), *пулемет* (скорость, стрельба, оружие, автомат, пуля), *разведка* (обследование, цель, действие, войско, группа, подразделение, дозор, получение, све́дение, противник, занятие, местность), *ружье* (рука, огонь, стрельба, оружие, длина, ствол), *тревога* (беспокойство, волнение, ожидание, опасность, неизвестность, сигнал, наступление), *чин* (служба, разряд, воинство, гражданин, чиновник, ведомство), *шеренга* (войско, строй, человек, линия), *щит* (предмет, рука, вооружение, круг, прямоугольник, плоскость, предохранение, удар), *автомат* (аппарат, устройство, включение, выполнение, самостоятельность, ряд, задание, операция, стрельба, оружие), *выстрел* (взрыв, заряд, канал, ствол, оружие, выброс, пуля, снаряд, дальность, звук), *сражение* (крупность, столкновение, бой, войско, армия), *танк* (броня, бой, машина, гусеница, ход, специальность, оборудование, резервуар, перевозка, хранение, жидкость), *фронт* (войско, строй, шеренга, обращение, противник, сторона, бой, расположение, район, действие, группа, армия), *армия* (вооружение, сила, оперативность, государство, войско, объединение, корпус, дивизия,), *артиллерия* (огонь, стрельба, орудие, род, войско, вооружение), *герой* (человек, храбрость, доблесть, совершение, подвиг, действие, лицо, литература, произведение), *затвор* (запор, устройство, механизм, машина, сооружение, оружие, место, жизнь, затворник), *корпус* (туловище, человек, животное, обособление, здание, ряд, крупность, войско, соединение, дивизия), *нападение* (бросок, вражда, намерение, начало, действие, противник, встреча), *пуля* (патрон, малость, снаряд, стрельба, ружье, винтовка, пулемет, револьвер), *заряд* (количество, взрыв, вещество, необходимость, содержание, устройство, соответствие, специальность, электричество, тело), *снаряд* (оболочка, взрыв, вещество, полет, выстрел), *бой* (вооружение, столкновение, сражение, разбиение), *война* (вооружение, борьба, государство), *стрельба* (произведение, выстрел, убийство, огонь, оружие, боль), *противник* (противодействие, враждебность, отношение, враг, соперник, войско, неприятель), *подразделение* (часть, раздел, вхождение, состав, крупность, войско, единица), *оружие* (средство, приспособление, техника, пригодность, нападение, защита), *укрепление* (делание, крепость, создание, оборона, сооружение, снабжение, средство), *окружение* (полнота, изоляция, группа, противник, остаток, войско, цель, уничтожение), *войско* (вооружение, сила, государство, часть), *богатырь* (герой, былина, совершение, воин, подвиг, человек, величина, сила, стойкость, отвага), *броня* (металл, одежда, воин, защита, туловище, прочность, воинство,

судно, танк, машина), *взвод* (приспособление, удар, механизм, курок, окружение, служение, упор, подъем, подразделение, рота, батарея, эскадрон), *дозор* (обход, осмотр, разведка, группа, войско).

Часть тела (44 слова): *остов* (внутренность, опора, часть, предмет, укрепление, костяк), *скелет* (совокупность, твердость, образование, состав, опора, остов, тело, человек, животное, каркас), *костяк* (скелет, основа, часть, организация), *крестец* (место, соединение, кость, таз, позвоночник), *позвоночник* (скелет, ось, образование, цепь, спина, заключение, канал, мозг), *позвонок* (обособление, кость, хрящ, элемент, позвоночник), *предплечье* (часть, рука, кисть, локоть), *колено* (сустав, соединение, бедро, голень, обособление, звено, отрезок, состав, изгиб, ход, разветвление, род, поколение, родословная), *лапа* (стопа, нога, животное, рука, человек, ветвь, хвоя, дерево), *стопа* (низ, часть, нога, щиколотка), *бедро* (часть, нога, сустав, таз, колено), *голень* (часть, нога, колено, стопа), *локоть* (место, сгибание, рука, соединение, плечо, кость), *плечо* (часть, туловище, промежуток, шея, рука), *щиколотка* (лодыжка), *ступня* (стопа, низ, часть, подошва), *лодыжка* (выступ, пара, сторона, сочленение, кость, стопа, щиколотка), *крыло* (о́рган, полет, птица, насекомое, аэродинамика, поверхность, аппарат, создание, подъем, сила, вращение, лопасть, колесо, ветер, двигатель), *конечность* (рука, нога, животное, обособление, о́рган, движение), *палец* (обособление, подвижность, конечность, кисть, стопа, лапа, человек, животное), *перо* (рог, образование, кожа, птица, стержень, пух, полость, отросток), *череп* (кость, образование, твердость, основа, голова, позвонок), *зуб* (кость, образование, о́рган, рот, схватывание, откусывание, жевание, пища),*хрящ* (позвонок, животное, человек, соединение, ткань, отличие, плотность, упругость), *губа* (подвижность, кожа, мышца, складка, образование, край, рот), *ухо* (о́рган , слух), *глотка* (участок, пищеварение, канал, соединение, полость, рот, пищевод, горло), *пасть* (зев, рот, зверь, рыба), *челюсть* (кость, основа, центр, низ, отдел, лицо, человек, о́рган, захват, пища, животное), *желудок* (о́рган, пищеварение, отдел, канал, следование, пищевод), *мозг* (о́рган, центр, нерв, система, вещество, образование, заполнение, череп), *сердце* (раздражение, гнев, центр, о́рган, кровь, несение, система, мышца, мешок, символ, душа, переживание, чувство), *спина* (часть, туловище, промежуток, шея, крестец), *зад* (часть, тело, противоположность, перёд, ягодица), *туловище* (тело, человек, животное, голова, конечность), *глаз* (о́рган, зрение, присмотр, надзор), *железа* (о́рган, выработка, выделение, гормон, вещество, обеспечение, жизнь, деятельность, организм), *мышца* (о́рган, движение, человек, животное, состав, ткань, способность, сокращение), *рот* (полость, верх, низ, челюсть, отверстие, лицо, очертание, разрез, губа, животное, пасть, едок, иждивенец), *язык* (о́рган, полость, рот, явление, вкус, человек, образование, звук, речь), *нога* (низ, конечность, человек), *голова* (часть, тело, человек, животное, состав, череп, коробка, лицо, морда), *рука* (верх, конечность, человек, плечо, палец), *запястье* (часть, кисть, рука, промежуток, предплечье, палец), *лицо* (перёд, часть, голова, человек, индивидуальность, облик), *пищевод* (отдел,

пищеварение, система, трубка, мышца, óрган, соединение, глотка, желудок, пища).

Биота (40): *клетчатка* (состав, растение, клетка, часть, оболочка, целлюлоза, волокно, соединение, ткань, организм, человек, животное), *лес* (площадь, земля, заросль, дерево, сруб, материал), *листва* (лист, растение, дерево, куст, покров), *морда* (перёд, часть, голова, животное, лицо), *мякоть* (мягкость, часть, тело, животное, человек, плод, клубень), *животное* (жизнь, организм, существо, обладание, способность, движение, питание, соединение, грубость, глупость, человек), *венчик* (часть, цветок, состав, обособление, сращивание, лепесток), *спора* (зачаток, растение, организм, гриб, размножение), *гриб* (обособление, организм, размножение, спора, нарост, тело, дерево) *клубень* (утолщение, корень, стебель, растение), *колос* (вид, соцветие, растение, злак, плетение, зерно), *колючка* (шип, острие), *паутина* (сеть, тонкость, нить, получение, выделение, паук, клейкость, быстрота, затвердение, жидкость), *рептилия* (пресмыкающееся), *заросль* (частота, кустарник), *лепесток* (тонкость, яркость, окраска, подобие, лист, пластинка, состав, венчик, цветок), *личинка* (земля, вода, рыба, животное, стадия, развитие, рождение, выход, яйцо, превращение, взросление, особь, куколка), *почка* (предшествование, развитие, побег, растение, начало, цветок, лист, вырост), *пресмыкающееся* (класс, позвонок, животное, передвижение, ползание, волочение, брюхо, земля, рептилия), *пыльца* (мужчина, пол, клетка, растение, нахождение, тычинка), *злак* (растение, стебель, полость, колено, соломина, мелкость, цветок, колос, метелка), *куст* (растение, дерево, вид, отсутствие, ствол), *млекопитающее* (животное, кормление, детеныш, молоко), *овощ* (огород, плод, зелень, употребление, пища), *пестик* (женщина, óрган, цветок, образование, плод, опыление, следствие), *тычинка* (мужчина, óрган, цветок, содержание. пыльца), *гусеница* (личинка, бабочка, червь, образ, множество, нога, ширина, цепь), *зародыш* (организм, начало, ступень, развитие, жизнь, счет, мать, питание, вещество, яйцо), *зерно* (мукá, плод, семя, злак, хлеб, малость, предмет, частица, зародыш), *ветка* (малость, бок, отросток, побег, дерево, кустарник), *кустарник* (заросль, куст, семейство, растение), *семя* (óрган, размножение, растение, зерно, предназначение, посев, зародыш, источник, сперма), *трава* (растение, мягкость, зелень, стебель), *птица* (покрытие, перо, пух, позвонок, животное, крыло, пара, конечность, клюв), *почва* (верх, слой, земля, кора, основание, опора), *стебель* (основа, часть, трава, растение, корень, вершина, цилиндр, форма), *природа* (существование, вселенная, мир, место, внешность, город, основа, свойство, сущность), *дерево* (растение, твердость, ствол, отход, ветвь, образование, крона, древесина), *бутон* (почка, цветок), *завязь* (расширение, часть, пестик, цветок, опыление, образование, плод).

Растение (7): *мак* (трава, растение, длина, стебель, крупность, краснота, цветок), *камелия* (украшение, растение, чай, зелень, вечность, лист, крупность, цветок, красота), *конопля* (трава, растение, семейство, тутовник, стебель, изготовление, пенька, семя, добыча, масло), *земляника* (трава, растение, дух, сладость, ягода, краснота, цвет), *лилия* (луковица, растение, крупность, красота,

прямая, стебель, цветок, вид, колокол), *лук* (огород, растение, семейство, лилия, острота, вкус, луковица, еда, трубка, лист, рука, оружие, метание, стрела, вид, пружина, дуга, натяжение, тетива), *виноград* (юг, лиана, растение, сладость, ягода, пища, изготовление, вино).

Зверь (1): *куница* (малость, хищник, зверь, ценность, мех).

Насекомое (4): *оса* (жало, полет, насекомое, перепонка, крыло, хищник), *паук* (хищник, насекомое, плетение, паутина), *шелкопряд* (бабочка, гусеница, витье, кокон, изготовление, шелк), *пчела* (жало, полет, насекомое, перепонка, крыло, общество, переработка, нектар).

Пресмыкающееся (2): *уж* (змея, вода, плавание, отсутствие, яд, полоз), *полоз* (гладкость, скольжение, сгибание, пластина, брус, змея, семейство, уж).

Религия (20): *икона* (изображение, бог, святой, портрет, образ, краска), *причащение* (христианство, таинство, принятие, причастие), *старообрядчество* (религия, течение, возникновение, следствие, раскол, стремление, сохранение, старина, церковь, правило), *храм* (здание, бог, служба, церковь), *таинство* (церковь, обряд, предназначение, приобщение, верующий, благодать), *творец* (человек, создатель, творчество, бог, мир), *православие* (основа, направление, христианство, вероисповедание, церковь, восток), *культ* (религия, служение, божество, связь, мистика, обряд), *священник* (служитель, культ, церковь, православие, сан, промежуток, дьякон, епископ), *суеверие* (вера, сверхъестественность, примета, тайна), *христианство* (религия, основание, положение, культ, миф), *приход* (поступление, сумма, деньги, товар, бухгалтерия, запись, книга, верующий, обслуживание, церковь), *бог* (религия, верховенство, всемогущество, существо, управление, мир, христианство, триединство, творец, начало, всеобщность, предмет, поклонение, обожание), *церковь* (объединение, последователь, религия, организация, ве'дение, жизнь, культ, православие, храм), *религия* (форма, общество, сознание, совокупность, духовность, представление, основание, вера, превышение, естественность, сила, существо, предмет, поклонение), *дьякон* (служитель, культ, православие, церковь, помощь, священник, совершение, служба), *епископ* (лицо, высота, священник, сан), *божество* (бог, предмет, обожание, поклонение), *верующий* (человек, вера, бог, религия, предписание, предрассудок, исполнение, церковь, обряд), *чудо* (сверхъестественность, бог, сила).

Смерть (10): *покойник* (смерть, человек, мертвец), *привидение* (сказка, мистика, представление, призрак, мертвец, воображение, существо), *повешение* (расположение, подвес, веревка, смерть, казнь, виселица), *гибель* (уничтожение, разрушение, смерть), *истребление* (уничтожение, гибель), *похороны* (обряд, закапывание, мертвец, земля, кремация), *мертвец* (смерть, человек, покойник), *смерть* (прекращение, жизнедеятельность, организм), *виселица* (пара, столб, перекладина, сооружение, смерть, казнь, повешение), *плоть* (тело, жизнь, смерть, человек, животное, растение).

Геология (3): *геология* (комплекс, наука, строение, состав, земля, польза, ископаемое, история, метод, поиск), *отложение* (гора, порода, образование,

результат, осадок, вода), *ископаемое* (добыча, недра, земля, существование, древность, геология, эпоха, нахождение, отложение, кора).

Математика (34): *интеграл* (математика, величина, рассмотрение, сумма, бесконечность, малость, часть), *касательная* (касание, прямая, общность, точка, кривая, разделение, плоскость, единственность, сторона), *логарифм* (показатель, степень, необходимость, возведение, данность, число, основание, получение), *минус* (математика, знак, отрицательность, вычитание, число), *множитель* (число, показатель, количество, повторение, слагаемое, получение, произведение), *производная* (образование, величина, форма, разность, приращение, отношение, предел, аргумент, функция, слово), *радиус* (расстояние, точка, окружность, центр, равенство, множество, нахождение), *тождество* (полнота, сходство, совпадение, математика, равенство, число, значение, величина), *формула* (краткость, точность, слово, выражение, определение, комбинация, математика, утверждение, символ, запись, химия, состав, строение, вещество), *частное* (число, получение, деление), *дифференциал* (математика, произвольность, бесконечность, малость, приращение, вариация, величина), *куб* (правильность, многогранник, грань, математика, число, показатель, степень, тройка, мера, объем, сосуд, перегонка, кипячение, жидкость), *уравнение* (математика, равенство, неизвестность, величина), *перпендикуляр* (линия, составление, угол, прямая, геометрия, фигура), *параллельность* (целое, протяжение, равенство, расстояние, сопоставление, совпадение, сопутствие, безразличие), *слагаемое* (число, сложение), *умножение* (арифметика, действие, повторение, число, слагаемое, единица, нахождение, данность, множитель), *арифметика* (учение, число, обозначение, цифра, действие), *грань* (плоскость, граница, образование, часть, поверхность, тело, ломаная), *треугольник* (геометрия, фигура, тройка, угол, многоугольник), *деление* (арифметика, действие, количество, содержание, число), *ломаная* (линия, соединение, угол, прямая), *окружность* (множество, точка, нахождение, одинаковость, расстояние, радиус), *функция* (явление, зависимость, изменение, мера, закон, значение, соответствие, аргумент, вариация, величина), *цифра* (знак, обозначение, число), *геометрия* (отдел, математика, изучение, пространство, форма, закон, измерение), *математика* (совокупность, наука, изучение, величина, количество, отношение, пространство, форма), *шар* (геометрия, тело, образование, вращение, круг, диаметр, предмет), *целое* (единство, величина), *степень* (мера, сравнение, величина, разряд, ступень, звание, повторение, умножение, число), *число* (понятие, количество, величина, помощь, произведение, счет), *длина* (протяжение, направление, пара, край, точка, линия, плоскость, тело, положение, максимум, расстояние), *единица* (цифра, изображение, число), *множество* (количество, совокупность, предмет).

Физика (22): *амплитуда* (размах, колебание, максимум, отклонение, равновесие, тело, положение), *нуклон* (элементарность, частица, ядро, состав, атом), *позитрон* (элементарность, частица, положительность, заряд, равенство, противоположность, электрон), *электрон* (элементарность, частица, несение,

отрицательность, заряд, получение, трение, янтарь), *нейтрон* (электричество, нейтральность, элементарность, частица, масса, равенство, протон), *протон* (элементарность, частица, вхождение, состав, атом, ядро, положительность, заряд), *проводимость* (способность, пропуск, электричество, ток), *потенциал* (физика, величина, характер, сила, поле, данность, точка), *излучение* (испускание, луч, поток, частица, элементарность), *луч* (узость, полоса, свет, направление, поток, частица, энергия, часть, прямая, линия), *ток* (поток, движение, масса, жидкость, воздух, нерв, энергия, человек, направление, электричество, заряд, проводник, место, птица, ритуал, пение, специальность, оборудование, площадь, хранение, обработка, зерно), *напряжение* (сосредоточение, сила, внимание, осуществление, величина, давление, растяжение, единица, площадь, разность, потенциал, электричество), *скорость* (степень, быстрота, движение, распространение, действие, физика, величина, равенство, расстояние, единица, время), *энергия* (основа, свойство, материя, мера, движение, способность, производство, работа), *электричество* (форма, энергия, обусловленность, движение, частица, материя, электрон, позитрон, протон, освещение), *физика* (основа, наука, естествознание, свойство, строение, материя, форма, движение, изменение, общее, закономерность, явление, природа), *масса* (основа, физика, характеристика, материя, определение, инертность, гравитация, свойство), *материя* (объективность, реальность, существование, независимость, человек, сознание, предмет, речь, разговор, ткань, вещество, состав, физика, тело), *жидкость* (вещество, обладание, свойство, течение, принятие, форма, сосуд), *газ* (шелк, прозрачность, ткань, вещество, состояние, частица, движение, свобода, распространение, полнота, пространство, заполнение, равномерность), *звук* (быстрота, колебание, движение, частица, воздух, среда, восприятие, слух, элемент, человек, речь, произношение), *вещество* (качество, проявление, материя, физика, тело, состав).

Химия (7): *углекислота* (химия, соединение, кислород, углерод), *щелочь* (едкость, химия, соединение, кислота, окрашивание, лакмус, бумага, синева, цвет), *окисление* (процесс, химия, реакция, кислород), *кислота* (химия, соединение, окрашивание, синева, лакмус, краснота, цвет, вкус), *соль* (белизна, кристалл, вещество, растворение, вода, острота, вкус, химия, соединение, металл, кислота, остаток), *атом* (мелкость, частица, химия, элемент, ядро, электрон), *кислород* (химия, элемент, газ, вхождение, состав, воздух, необходимость, горение, дыхание).

Медицина (8): *бинт* (длина, узость, лента, марля, хло́пок, ткань, лечение, повязка), *хирургия* (раздел, медицина, болезнь, операция, метод, лечение), *хирург* (врач, специалист, хирургия), *заболевание* (болезнь, возникновение), *лекарство* (лечение, средство), *медицина* (совокупность, наука, здоровье, болезнь, лечение, предупреждение, практика, деятельность, сохранение, укрепление, человек), *лечение* (применение, медицина, средство, избавление, болезнь, принятие, мера, прекращение), *болезнь* (расстройство, здоровье, нарушение, деятельность, организм).

Искусство (4): *зодчество* (архитектура), *архитектура* (искусство, проектирование, строение, здание, сооружение, зодчество, стиль, постройка), *скульптор* (художник, изготовление, произведение, скульптура), *скульптура* (произведение, искусство, создание, объем, художественность, изображение).

Музыка (13): *напев* (мелодия), *обертон* (звук, произведение, колебание, струна, частота, кратность, основа, тон, придание, обособление, тембр), *певец* (человек, умение, пение, артист, вокал), *педаль* (нога, рычаг, машина, музыка, инструмент), *вокал* (певец, искусство), *тембр* (характер, окраска, звук, сообщение, обертон), *мелодия* (благозвучие, последовательность, звук, образование, известность, музыка, единство, напев), *пение* (издание, голос, музыка, звук, исполнение, вокал, произведение, птица, свист, восхваление), *тон* (звук, определенность, высота, музыка, работа, сердце), *голос* (совокупность, звук, возникновение, результат, колебание, связка, мелодия, музыка, пьеса, мнение, высказывание, внутренность, побуждение, осознание, право, заявление), *инструмент* (рука, орудие, производство, работа, музыка, специальность, прибор, извлечение, звук), *музыка* (искусство, переживание, чувство, идея, выражение, ритм, интонация, организация, звук, произведение, исполнение, инструмент, мелодия), *благозвучие* (приятность, слух, сочетание, звук, музыка).

Поэзия (2): *поэзия* (слово, художественность, творчество, стихотворение), *стихотворение* (малость, поэзия, произведение).

Театр (9): *кинематограф* (театр, показ, экран, фильм), *кинотеатр* (место, просмотр, фильм, кино), *актер* (артист, исполнитель, роль, театр, представление, кино, телевидение), *пьеса* (драма, произведение, театр, представление, музыка), *публика* (лицо, нахождение, качество, зритель, слушатель, пассажир, общество), *кино* (фильм, кинотеатр, кинематограф, смех, событие, ситуация), *фильм* (тонкость, прозрачность, пленка, вид, лента, целлулоид, свет, чувствительность, слой, кино), *сцена* (специальность, площадка, происхождение, театр, представление, деятельность, обособление, часть, действие, эпизод, пьеса, повесть, роман), *театр* (искусство, изображение, драма, произведение, сцена, помещение, представление, война, действие, анатомирование, труп).

Цирк (6): *акробат* (артист, цирк, спортсмен, прыжок, равновесие), *реприза* (краткость, шутка, цирк, спектакль, представление), *фокус* (номер, цирк, представление, трюк, искусство, обман, зрение, внимание, помощь, ловкость, прием, точка, пересечение, преломление, луч, свет, падение, оптика, система, параллельность, пучок, объектив, четкость, изображение), *цирк* (вид, искусство, объединение, выступление, акробат, гимнаст, клоун, фокус, животное, дрессировщик), *дрессировщик* (артист, цирк, управление, животное), *клоун* (артист, цирк, реприза, маска, смех).

Семья и брак (14): *бракосочетание* (обряд, церемония, вступление, брак), *девочка* (ребенок, женщина, пол), *мальчик* (ребенок, мужчина, пол, слуга, подросток, дом, заведение), *любовник* (мужчина, нахождение, пол, связь, женщина, состояние, брак, влюбленность, человек, роль, театр, пьеса),

сладострастие (сила, влечение, удовлетворение, пол, чувство, наслаждение, страсть, увлечение), *сожительство* (совместность, жизнь), *супруг* (муж), *супружество* (брак, жизнь), *влюбленность* (испытание, страсть, любовь, влечение), *измена* (предательство, интерес, народ, переход, сторона, враг, нарушение, верность), *родственник* (человек, нахождение, родство), *муж* (мужчина, отношение, женщина, состояние, брак, супруг), *семья* (группа, жизнь, близость, родственник, совместность, объединение, человек), *брак* (семья, супружество, отношение, мужчина, женщина, сопровождение, обряд, бракосочетание).

Утварь (17): *веник* (связка, ветка, прут, подметание), *метла* (предмет, хозяйство, обиход, подметание, прикрепление, палка, пучок, прут), *щетка* (изделие, очистка, подметание, плоскость, колодка, частота, насадка, пук, щетина, жесткость, волос), *оглобля* (упряжь, круг, длина, жердь, укрепление, повозка, соединение, дуга), *ступица* (центр, часть, колесо, отверстие, насадка, ось, обод, соединение, спица, диск), *сундук* (тяжесть, пол, ящик, хранение, вещь, крышка, петля, замóк), *хомут* (часть, упряжь, шея, лошадь, круг, дерево, остов), *чашка* (малость, сосуд, округлость, форма, тарелка, подвес, коромысло, весы), *чугун* (плавление, руда, железо, примесь, углерод, хрупкость, сосуд, округлость, форма), *конфорка* (надставка, труба, самовар, подогревание, чайник, диск, закрытие, отверстие, кухня, плита), *лемех* (часть, плуг, резание, почва, отвал, передача), *поднос* (металл, плоскость, лист, изгиб, край, верх, перенос, посуда, подача, еда, стол), *узда* (часть, сбруя, ремень, удила, повод, надевание, голова, упряжь, животное), *замóк* (приспособление, запирание, ключ, затвор), *молоток* (металл, дерево, брусок, насадка, прямоугольник, рукоятка, служение, битье, удар), *плуг* (село, хозяйство, орудие, ширина, металл, лемех), *сбруя* (предмет, седло, лошадь), *упряжь* (совокупность, принадлежность, соединение, животное, повозка, сбруя).

Наиболее популярной темой в ядре является военная: слов, напрямую связанных с понятием «война» и вовлеченных в короткие циклы длин 2, 3, 4, в ядре оказалось 55. На втором месте идут понятия, связанные со строением человеческого тела, на третьем – с животным и растительным миром. После них наиболее востребованными являются точные науки – математика (34 слова) и физика (22 слова). В средней части ядра, разбитого на совокупность коротких циклов, находятся искусства, ремесла, ритуалы, а также часто употребляемые слова-связки, такие как «часть», «форма», «предмет» и т.п.

Всего в ядре можно выделить следующие 20 крупных тематических кластеров: 1 - война, 2 - человек, 3 - живая природа, 4 - наука, 5 - искусство, 6 - религия, 7 - семья, 8 - быт, 9 - мораль, 10 - техника, 11 - действие, 12 - жизнь, 13 - смерть, 14 - чувства, 15 - состояние, 16 - еда, 17 - политика, 18 - управление, 19 - окружающая среда, 20 - связующие термины.

Каждый из этих кластеров может быть, в свою очередь, разбит на уровни. Например, «война» включает в себя «способы ведения», «объекты нападения», «изделия», «боевой дух» и т.п. Кластер «человек» делится на «скелет», «органы», «мускулатура». Кластер «наука» содержит математику, физику,

биологию, медицину и т.д. Это кластеры второго уровня. При необходимости детализации смысла высказывания эти кластеры могут быть разбиты на кластеры третьего уровня: математика (алгебра, геометрия, …), физика (механика, термодинамика, электричество, …) и т.д.

4. Оценка близости смыслов высказываний

В работе [1] был предложен метод определения смысла предложения на основании приближения совокупности циклов как переполненного базиса. В настоящей работе рассматривается другой метод, основанный на более адекватном с практической точки зрения кластерном анализе близости высказываний. Опишем здесь соответствующую конструкцию.

Предположим, что в результате экспертного мнения слова сильно связной компоненты были разбиты на M_1 кластеров первого уровня, так что про каждое слово известно, какому смысловому кластеру оно принадлежит. Пример такого экспертного разбиения был приведен выше в п. 3. Насколько точно или адекватно проведено такое разбиение, для дальнейшего роли не играет, поскольку близость между словами и высказываниями будет определяться тем, с какими словами из каких кластеров данное слово связано циклами. Кластеры нужны только затем, чтобы обозначить темы, интересующие исследователя.

Далее определим вероятность, с которой смысл слова (высказывания) принадлежит тому или иному кластеру.

Пусть данное слово ядра A_i, где $i = 1, 2, \ldots, N = 3874$, связано циклами длины 2 с числом $n_j^{(2)}(i)$ слов из j-го тематического кластера. Если других циклов мы не рассматриваем, то в приближении равнозначности всех слов ядра вероятность $P_j(i)$ того, что смысл слова под номером i находится в кластере с номером j, равна

$$P_j(i) = \frac{n_j^{(2)}(i)}{\sum_{m=1}^{M_1} n_m^{(2)}(i)}.$$

Здесь вероятность трактуется как эмпирическая частота попадания слова, входящего в цикл, в заданный тематический кластер.

Если рассматриваются также и циклы длины 3, то в j-ый кластер попадают слова как из циклов длины 2, так и из циклов длины 3. Количество слов, попадающих в j-ый кластер и связанных циклом длины 3 со словом i, обозначим $n_j^{(3)}(i)$. Далее следует ввести предположение о весовой функции цикла в образовании смысла слова. Из предыдущего анализа следует, что с увеличением длины цикла число циклов, которых надо удерживать в памяти, резко возрастает, поэтому естественно ввести вес цикла данной длины, обратно пропорциональный соответствующему числу циклов. Положим

$$w_l = \frac{1}{n(l)},$$

где $n(l)$ есть число циклов длины l (см. табл. 1 и аппроксимацию после нее). Тогда для искомой вероятности получаем формулу

$$P_j(i) = \frac{\sum\limits_l w_l n_j^{(l)}(i)}{\sum\limits_s w_s \sum\limits_{m=1}^{M_1} n_m^{(s)}(i)}. \qquad (1)$$

Суммирование в (1) проводится в идеале по всем циклам ядра. Однако, памятуя об их фактически бесконечном количестве, на практике реализовать эту формулу можно лишь приближенно, ограничиваясь относительно короткими циклами. Тем не менее, поскольку все слова ядра должны быть охвачены циклическими связями, учитывать надо циклы вплоть до длины $L = 13$, число которых оценивается в 60 млрд. Однако вклад этих последних в итоговую вероятность (1) будет пренебрежимо мал.

В результате с каждым словом ядра будет связан M_1-мерный вектор вероятностей $P_j(i)$, для каждого i нормированный на единицу:

$$\sum_{j=1}^{M_1} P_j(i) = 1.$$

Если нас интересует более детальная смысловая информация, то каждый кластер первого уровня экспертно разбивается на подклассы второго уровня, число которых обозначим $M_2(j)$. Вводя аналогично предыдущему число заполнения $n_{jk}^{(l)}(i)$ k-го подкласса j-го кластера словами из циклов длины l, получаем выражение для двухуровневой вероятности

$$P_{jk}(i) = \frac{\sum\limits_l w_l n_{jk}^{(l)}(i)}{\sum\limits_s w_s \sum\limits_{m=1}^{M_1} \sum\limits_{r=1}^{M_2(m)} n_{mr}^{(s)}(i)}. \qquad (2)$$

Предполагая, что третьего уровня детализации смысла достаточно, приведем формулу для трехуровневой вероятности, связанной с данным словом ядра словаря:

$$P_{jkq}(i) = \frac{\sum\limits_l w_l n_{jkq}^{(l)}(i)}{\sum\limits_s w_s \sum\limits_{m=1}^{M_1} \sum\limits_{r=1}^{M_2(m)} \sum\limits_{t=1}^{M_3(m,r)} n_{mrt}^{(s)}(i)}. \qquad (3)$$

Близость между словами и высказываниями естественно оценивать как близость между соответствующими векторами (1), (2) или (3) в норме $L1$, например:

$$\rho(A,B)=\sum_{j=1}^{M_1}\left|P_j(A)-P_j(B)\right|. \qquad (4)$$

Если «*A*» – это высказывание, состоящее из *A* слов – существительных, то

$$P_j(A)=\frac{1}{A}\sum_{a=1}^{A}P_j(a), \qquad (5)$$

где a есть номер слова ядра, входящего в данное высказывание. Той же формулой (5) описывается вероятность для слов из оболочки или канала словаря, поскольку они известным путем связаны с ядром.

Отметим, что с высказыванием можно попробовать связать отдельное слово, которое наиболее близко в смысле формулы (4) подходит к данному высказыванию. Такое слово или совокупность слов может представлять собой краткое изложение смысла более длинного предложения или совокупности предложений.

Таким образом, концепция «циклического смысла» слова или высказывания может быть эффективно использована для проведения оценки семантической близости различных высказываний.

Литература

1. Орлов Ю.Н., Парфенова Ю.А. Анализ структуры онтологического графа толкового словаря // Препринты ИПМ им. М.В. Келдыша. 2017. № 31. 25 с. URL: http://library.keldysh.ru/preprint.asp?id=2017-31
2. Orlov Y.N., Parfenova Y.A. Structure analysis of an explanatory dictionary ontological graph // Lecture Notes in Computer Science. 2017. V. 10531. P. 408-419.
3. Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. – М.: Русский язык, 1989.