

Ассоциативная сеть понятий, образующих запросы к интернету

An associative network of concepts that enter to internet queries

Большаков И. А. (bolshakov34@mail.ru),
Независимый исследователь, Москва

Большакова Е. И. (bolsh@cs.msu.ru),
Московский государственный университет, Москва

Гельбух А. Ф. (gelbukh@gelbukh.com),
Национальный политехнический институт, Мехико, Мексика

В базе пользовательских запросов поисковиков Google и Яндекс выявлена обширная совокупность сочиненных пар существительных и на их основе построена и описана ассоциативная сеть понятий, из которых часто формируются русскоязычные запросы к Интернету. Показано, что выявленные пары существительных предствительно входят и в текстовые массивы Интернета. Исследована полученная ассоциативная сеть.

1. Введение

Уже более 10 лет интернетовские поисковые машины типа Google накапливают базу пользовательских запросов (БПЗ) для своей внутренней надобности — совершенствования поискового аппарата. Запрос может состоять из одного или нескольких слов, быть грамматически связанным (*производство дверей*) или несвязанным (*кирпич цена*), орфографически безупречным или содержащим ошибку.

Пару лет назад Google позволил пользователям оперативно общаться с этой БПЗ. В процессе посимвольного ввода пользователем запроса поисковик выводит и модифицирует меню из упорядоченных подсказок, начальные символы которых совпадают с уже введенной частью запроса. Часто можно не вводить свой запрос целиком, а лишь выбрать одну из строк образующегося меню.

Обнаружилось, что при запросе с более чем одним понятием пользователи, не прибегая к средствам расширенного поиска с формальными логическими связками, часто соединяют понятия союзом *и*. Например, если беременная пользовательница хочет оценить для себя опасность аллергии, она вводит запрос *беременность и _*, (*_* означает пробел), а затем выбирает в образовавшемся меню строку *беременность и аллергия*. При этом Google сообщает, что таких запросов уже было 515 тыс., а это зна-

чит, что запрос привычен для поисковика и на него есть релевантные ответы. Доступ к БПЗ сейчас есть и у русских поисковиков (Яндекс, Рамблер). При этом русская БПЗ Яндекса оказалась побогаче, чем у Google и Рамблера.

Тем самым современные поисковики предоставляют прикладному лингвисту среди многого прочего тысячи сочиненных пар указанного вида. Возникает возможность создать или резко пополнить коллекцию сочиненных именных пар. Это не только интересно с языковой точки зрения, но и позволяет создать ассоциативную сеть понятий, которыми оперирует русскоязычный пользователь Интернета. В части ассоциаций здесь никаких методологических проблем не возникает, но в принадлежности некоторых сочиненных пар из БПЗ русскому языку в целом можно и усомниться. Действительно, можно ли считать принадлежащими русскому языку такие пары, как *ноги и характер* (2,49 млн. запросов к Google, январь 2010 г.) или *ангина и керосин* (12,1 тыс. запросов)?

Данная работа преследует следующие цели:

- Описать коллекцию сочиненных именных пар, входящую в словарь КроссЛексика [1], после существенного пополнения ее данными из Google и Яндекса;
- Дать приближенную интерпретацию ряда характерных запросов из сочиненной пары, по-

казав на примерах несводимость возникающих ассоциаций к семантическим связям типа WordNet [3];

- На основе статистических данных показать, что новые пары понятий встречаются и на сайтах Интернета, а потому могут считаться принадлежащими русскому языку в целом;
- Построить из компонентов сочиненных пар ассоциативную сеть понятий, которыми оперирует русскоязычный пользователь в запросах к Интернету;
- Проанализировать построенную сеть, выявив понятия с максимальным количеством ассоциативных связей, вычленив и описав связанные компоненты сети и др.

2. Пополненная коллекция сочиненных пар

В коллекции сочиненных пар словаря КроссЛексика [1, 2] до привлечения содержимого БПЗ было 4,1 тыс. пар, бытующих в текстах и подтвержденных сайтами Интернета. Из них субстантивных пар было 2,8 тыс., т. е. 68%. Примерно в 10 процентах случаев компоненты субстантивных пар были не одиночными существительными, а группами типа *авторские права*, *автоматическое оружие* или *железные дороги*.

Семантически компоненты пар соотносились по одному из следующих вариантов:

- когипонимы в некой родовидовой иерархии (*руки и ноги*, *аксиомы и теоремы*, *труд и капитал*, *акушерство и гинекология*);
- синонимы, квазисинонимы и повторы (*траур и скорбь*, *горести и несчастья*, *тысячи и тысячи*);
- антонимы, квазиантонимы, противоположные понятия и конверсивы (*бедные и богатые*, *актив и пассив*, *Бог и дьявол*, *купля и продажа*, *действие и противодействие*);
- парные названия и исторически связанные имена (*Босния и Герцеговина*, *Адам и Ева*).

Лишь изредка компоненты оказывались соучастниками некоторой ситуации (*писатель и читатели*, *закон и порядок*, *кожа и косметика*) или понятиями, связанными причинно-следственными связями (*война и разруха*, *преступление и наказание*, *штормы и наводнения*).

Для целей данной работы мы брали пары не только из Google, но и из Яндексa, поскольку статистического порога извлечения новых пар из БПЗ не устанавливалось. В первом туре пополнения в качестве первого компонента пары брались субстантивы *S*, уже содержащиеся в коллекции в качестве первого либо второго компонента. Вручную вводились запросы вида *S_u_*. В окне поисковика возникали различные пары, частично новые для коллек-

ции. Последними и велось пополнение. При этом в качестве второго компонента в новых парах часто появлялись и новые понятия. Они брались началом аналогичных запросов во втором туре, что выявляло новые компоненты на второй позиции в паре. В третьем туре новые вторые компоненты брались первыми компонентами очередных запросов и т. д. После семи туров пополнения общее количество пар достигло 16 тыс., а число субстантивных пар превысило 14 тыс. (более 90 % всей коллекции), увеличившись в 5,6 раз.

Чтобы обеспечить смысловую самостоятельность обоих компонентов пар, мы не брали довольно частые пары с анафорической отсылкой *его/ее/их* во втором компоненте (например, *кризис и его последствия*), а также пары типа *газеты и другие/иные СМИ*. Были отброшены повторы. Замеченные редкие синонимы, иногда — абсолютные (*Белоруссия = Беларусь*, *кормление грудью = грудное вскармливание*, *зарплата = заработная плата*) не объединялись.

Смысл сочинения компонентов-понятий в парах стал после пополнений много разнообразнее. Охарактеризуем это на примерах. Так, множество пар *X* и *цены* эквивалентно по смыслу *цены(X)?*, т. е. пользователь просто заменил в запросе подчинительную связь сочинительной. Это не касается пар *цены и комплектация/наличие/скидки/ценообразование*, где пользователь интересуется разными параметрами некоторого объекта, непосредственно парой не отражаемого (см. табл. 1).

Более сложен смысл сочинения у образцов запросов к Google, представленных в табл. 1. Например, для большинства пар, включающих *Y = беременность / здоровье*, семантическим представлением запроса является *влияние(X,Y)?* первого аргумента на второй, например, *влияние(простуда, беременность)?*. Для пар, включающих *СМИ*, запрос примерно эквивалентен симметричному предикату *взаимодействие(X,Y)?*, например, *взаимодействие(СМИ, власть)?* В парах, где *X = йога*, а *Y = православие / христианство / буддизм*, замечаем симметричный предикат *совместимость(йога,Y)?* Осмысление пары *ангина и керосин* позволяет предположить структуру уже с двумя предикатами: *эффективность(лечение(ангина, керосин))?* Подобные толкования разного уровня сложности могут быть предложены и для всех иных запросов.

Скрытая семантика запросов, как видим, разнообразна, а компоненты пар связаны неизвестным предикатом, и быть может, не одним. Эти предикаты может предположить лингвист, но алгоритмически они в общем случае не восстанавливаемы. Поэтому поиск простых и коротких семантических связей между компонентами пары {*X*, *Y*} средствами, например, WordNet [3], который оперирует синонимами, антонимами, гипонимами / гиперонимами и меронимами / холонимами, в общем случае без-

Таблица 1. Статистика образцов запросов и ответов

(VQ — число запросов, VS — число прямых ответов, VF — число косвенных ответов, все в тысячах)

п/п	Сочиненная пара	VQ	VS	VF
1	беременность и роды	1470,0	1380,0	1720,0
2	беременность и простуда	219,0	249,0	263,0
3	беременность и компьютер	784,0	99,2	834,0
4	беременность и месячные	271,0	201,0	251,0
5	беременность и курение	494,0	52,0	499,0
6	беременность и питание	1450,0	37,8	1470,0
7	беременность и грипп	460,0	258,0	593,0
8	беременность и молочница	171,0	125,0	163,0
9	здоровье и красота	99700,0	2110,0	144000,0
10	здоровье и материнство	108,0	118,0	195,0
11	здоровье и спорт	315000,0	173,0	261000,0
12	здоровье и комфорт	915,0	178,0	926,0
13	здоровье и здоровый образ жизни	1960,0	81,6	1180,0
14	здоровье и долголетие	243000,0	40,1	310000,0
15	здоровье и окружающая среда	558,0	121,0	426,0
16	здоровье и питание	266000,0	455,0	495000,0
17	здоровье и компьютер	158000,0	21,2	269000,0
18	СМИ и периодика	249,0	124,0	202,0
19	СМИ и политика	9670,0	54,9	10400,0
20	СМИ и власть	3690,0	5,8	3570,0
21	СМИ и культура речи	561,0	13,3	315,0
22	СМИ и общество	8930,0	561,0	10300,0
23	СМИ и общественное мнение	2180,0	39,4	1350,0
24	СМИ и PR	2860,0	364,0	2840,0
25	СМИ и государство	14100,0	298,0	8520,0
26	СМИ и выборы	2300,0	116,0	986,0
27	СМИ и культура	8550,0	203,0	9700,0
28	ангина и беременность	64,8	34,0	67,1
29	ангина и антибиотики	181,0	1,2	139,0
30	ангина и кашель	584,0	42,1	1110,0
31	ангина и гомеопатия	50,6	0,0	18,9
32	ангина и грудное вскармливание	240,0	2,2	767,0
33	ангина и керосин	12,0	0,0	5,8
34	ангина и сердце	5010,0	6,3	7300,0
35	ноги и характер	2490,0	198	2370,0
36	алгоритмы и структуры данных	359,0	10,2	252,0
37	алгоритмы и исполнители	80,6	10,8	69,2
38	алгоритмы и программы	720,0	160,0	592,0
39	алгоритмы и процессоры	63,1	46,2	170,0
40	алгоритмы и блок-схемы	96,8	1,5	94,0
41	аккумуляторы и зарядные устройства	3230,0	97,1	2320,0
42	аккумуляторы и цены	2500,0	0,0	2290,0
43	йога и питание	20800,0	1200,0	32300,0
44	йога и православие	197,0	23,1	138,0
45	йога и беременность	290,0	97,1	372,0
46	йога и здоровье	45200,0	244,0	48200,0
47	йога и похудение	253,0	1,2	143,0
48	йога и алкоголь	157,0	11,1	151,0
49	йога и христианство	186,0	9,3	194,0
50	йога и сколиоз	10,1	1,1	16,3

надежен. Но для десятков, сотен тысяч или даже миллионов пользователей Интернета эти компоненты очевидно связаны по смыслу, а потому введение в оборот компьютерной лингвистики отдельной семантической подсистемы таких связей представляется оправданным.

Специфичность сочинения в запросах позволяет усомниться в наличии некоторых подобных пар в обычных текстах языка. Для прояснения ситуации мы произвольно взяли из Google 50 пар со статистикой БПЗ (столбец VQ табл. 1), а также измерили число страниц в основной базе данных Google, которые включают те же пары буквально (т. е. контактно и в номинативе, столбец VS) или на произвольном расстоянии внутри абзаца и произвольным падежом (столбец VF). Все числа измерялись в тысячах. Будем называть столбец VS статистикой прямых ответов, а столбец VF — косвенных ответов. За исключением трех случаев (6%), сайты Google содержат прямые ответы, и среднее отношение числа прямых ответов к числу запросов составляет $k_{sq} = 0,19$. Косвенные ответы встречаются на сайтах много чаще, со средним коэффициентом пропорциональности $k_{fq} = 1,1$.

Поскольку конкретное значение коэффициента k_{sq} резко меняется от пары к паре, вычислим еще косинусы углов между векторами VQ, VS, VF с координатами $i = 1 \dots 50$. Косинус угла между VQ и VS дается формулой

$$\cos(VQ, VS) = \frac{\sum_i (VQ_i VS_i)}{\sqrt{\sum_i (VQ_i)^2 \sum_i (VS_i)^2}},$$

а для прочих пар — аналогичными формулами, что дало $\cos(VQ, VS) = 0,26$; $\cos(VQ, VF) = 0,95$; $\cos(VS, VF) = 0,27$. Иными словами, вектора запросов и косвенных ответов практически коллинеарны, но различие между векторами запросов и прямых ответов существенно. Поэтому запросы в виде пар могут считаться принадлежащими русскому языку лишь в своем большинстве.

3. Семантическая сеть запросов

После пополнения коллекции сочиненных пар был программно построен неориентированный граф с вершинами, помеченными различными понятиями. Ребра графа соединяют вершины с понятиями X, Y, если последние входят в пару X и Y и/или X и Y. Например, вершина ангины связана с вершинами антибиотиков, беременность, гомеопатия, грудное вскармливание, кашель, керосин, мороженое, прополис, сердце2, фарингит. Здесь и далее цифра в конце слова означает номер омонима.

Таблица 2. Степени D наиболее популярных понятий

D	Понятие	D	Понятие
302	беременность	31	ремонт
110	здоровье	29	армия
87	алкоголь	29	методы
87	цены	29	экономика
54	спорт	28	давление1
52	культура1	28	лечение
51	похудение	28	функции1
49	дети	27	безопасность
48	человек	27	власть
41	диабет	27	реклама
40	диета	27	экология
39	курение	26	структура1
39	любовь	25	философия
37	общество1	24	контроль
37	религия	24	Наука
37	Россия	24	пиво
36	право2	24	христианство
34	температура	23	Водка
34	характер2	23	государство
33	бизнес	23	деньги
33	дизайн	23	Интернет
32	кризис	23	искусство
32	развитие	23	православие
31	политика	23	прыщи

Полученный граф назовем ассоциативной сетью [4]. На март 2010 г. она включала 9,2 тыс. вершин-понятий с общим числом ребер-ассоциаций 25,2 тыс. (в среднем 2,73 ребер на вершину). В таблице 2 дана степень D (число ребер) наиболее популярных понятий. Два первых места с большим отрывом занимают беременность и здоровье. Далее устойчиво расположились алкоголь, цены, спорт, культура1, похудение, дети и человек. Значит, все это наиболее острые проблемы для пользователей. Алкоголь поднимается на одну позицию, если присоединить пиво, водка, алкогольные напитки и алкогольики. Но все, прямо связанное со здоровьем (беременность, здоровье, спорт, похудение, диабет, диета и др.), по совокупности всегда первенствует. Среди лидеров всегда остаются также цены, но они интересны далеко не для всех товаров и услуг, вошедших в сеть.

Таблица 3. Степени D популярных многословных понятий

D	Понятие
22	окружающая среда
20	щитовидная железа
16	кормление грудью

D	Понятие
14	лунный календарь
13	грудное вскармливание
12	социальная политика
12	характерные черты
11	государственное управление
11	группа крови
11	международное право
11	охрана окружающей среды
10	охрана природы
10	рынок труда
10	экономический рост
9	витамин С
9	глобальные проблемы
9	заработная плата
9	культура речи
9	Новый год
9	общественное мнение
9	социальный контроль
8	бронхиальная астма
8	зеленый чай
8	знаки зодиака
8	информационные технологии
8	образ жизни
8	оливковое масло
8	охрана труда
8	рыночная экономика
8	социальная справедливость
8	тепловые двигатели
7	валютный курс

Многословные понятия представлены в сети достаточно широко. В таблице 3 даны наиболее популярные из них. Интерес пользователей к окружающей среде и ее охране, щитовидной железе, кормлению грудью, группе крови, рынку труда, витамину С и заработной плате кажется естественным. Удивляет, однако, высокая популярность лунного календаря, выдающая интерес широкой массы пользователей к эзотерике. Отметим еще популярность понятий социальной политики, государственного управления, международного права, экономического роста, информационных технологий, тепловых двигателей и др., что, похоже, отражает интерес студентов разных специальностей к материалам для рефератов и курсовых, добываемым из Интернета.

Изучим теперь связность созданной ассоциативной сети: распадается ли она на изолированные компоненты связности (подсети), и если да, то каковы эти подсети? Путем программного обхода вершин были найдены все компоненты связности и рассчитаны их параметры (см. табл. 4). **Мощностью** графа называется число узлов в нем [4]; **диаметром** — са-

мая длинная из кратчайших цепей, связывающих какие-либо две вершины графа; **мостом** — ребро, разрыв которого увеличивает число компонентов связности графа; **точкой сочленения** — вершина, удаление которой ведет к увеличению числа компонентов связности; **степенью вершины** — число ребер, которым она принадлежит; **висячей вершиной** — вершина степени 1.

Всего в сети (на январь 2010 г.) было 870 связанных подсетей. Доминирующая подсеть содержала 5129 вершин. Ее тематика предельно широка, мы назвали ее общежитейским универсумом. Сюда вошли все понятия таблиц 2 и 3. С огромным отрывом далее шла подсеть о воспитательно-образовательной сфере с 21 вершиной. На третьем месте — две подсети по 13 вершин, отражающих детали домов и преступность соответственно. На предпоследнем по мощности месте оказалось 117 подсетей из трех вершин, на последнем месте — 654 подсети из двух вершин. Тематика подсетей из 2–5 вершин узка и многообразна, мы обозначили ее как <фрагменты>. Можно заметить, что диаметры часто включают мосты, а мосты включают многие точки сочленения.

Особый интерес вызывает доминирующая подсеть. Ее размеры столь велики, что программное (кубической сложности) вычисление ее диаметра оказалось неосуществимым даже за неделю счета. Но анализ показал, что 2836 вершин этой подсети (т. е. более половины всех вершин) являются висячими. Подсеть-доминанта похожа на ежа с громадным количеством торчащих иголок типа *абазуры*, *аббревиатуры*, *абитуриенты*, *абсурд*, *авансирование*, *автоаксессуары*, *автокосметика*, *автокредит*, *авторизация*, *авторитаризм*, *автохимия*. Почти все эти понятия редки в обиходе. Если отделить висячие вершины, оставшаяся более компактная и более связанная подсеть из 2208 вершин уже позволяет вычислить ее диаметр за ограниченное время (снабжен знаком + в таблице 4).

Полезно заметить, что пополнение сети систематически ведет к слиянию малых подсетей с доминантой. Так, появление ассоциации *ложки — столовые приборы* вызовет присоединение к доминанте подсети из восьми слов (см. таблицу 4), поскольку в доминанте уже содержится вершина *столовые приборы*. Эта тенденция не исключает появление небольших изолированных подсетей при будущих пополнениях, и до некоторого этапа они могут разрастаться отдельно, но потом сольются с доминантой. При всем этом всегда сохранится множество двухвершинных подсетей типа *Адам — Ева*. Для входящих в них понятий трудно вообразить иные ассоциации, выразимые сочиненными парами.

Образование всеподавляющей доминантной подсети внутри сформированной полной сети вполне коррелирует с результатами тех исследований по семантике слов средствами WordNet [5], которые основаны на теории случайных графов [6, 7].

Таблица 4. Связные подсети и их характеристики

Мощность	Подсетей	Длина диам.	Примеры диаметров	Примеры мостов	Примеры тчк. сочленения	Тематика
5129	1	14+	продавцы–покупатели–поставщики–закупки–снабжение–комплектация–цены–ламинанат–вода–ветер–снег–грозы–дожди;	гололедица–снег; комплектация–цены; питательные вещества–вода;	цены; снег; вода; водка;	общежитейский универсум
21	1	10	любители–профессионалы–дилетанты–специалисты–ЕГЭ–вузы–школы1–колледжи–лицей–гимназии;	<i>специалисты–ЕГЭ; ЕГЭ–вузы; лицеи–гимназии; ясли–детсады;</i>	вузы; ЕГЭ; детсады; институты; специалисты	воспитательно-образовательная сфера
13	2	7	фасад–кровля–фасады–кровли–крыша1–перекрытия–пустоты;	изоляция–кровли; кровли–фасады;	кровли; крыши; перекрытия;	(1)детали домов (2)преступность
11	3	6–8	диаметр–окружность–круг1–крест–шар–сфера1;	диаметр–окружность; круг1–крест;	крест; круг1; окружность; шар	(1)геометр. фигуры (2)фазы изменения (3)стройматериалы
10	3	5–7	мародерство–грабежи–убийства–теракты–диверсии;	диверсии–теракты; попытки–убийства;	грабежи; теракты; убийства	(1)преступные акты (2)выч.математика (3)лингвистика
9	2	3, 4	БМП–танки1–артиллерия–ПВО;	БМП–танки1; артиллерия–ПВО;	артиллерия танки1	(1)боевая техника (2)студенч.работы
8	7	3–6, 8	<i>перила–лестницы–лифты–подъемники; перчатки–шарфы–палантины;</i>	лестницы–лифты; лифты–подъемники; варежки–шарфы;	лестницы; лифты; шарфы	(1)лестницы–лифты (2)одежда (3)оружие (4)насекомые (5)термодинамика (6)деды–внуки (7)авторы–герои
7	4	3, 4	хранение файлов–хостинг–домен–рабочая группа;	хостинг–домен;	хостинг; домен	(1)сайты–домены (2)персон. черты (3)покрытия (4)волны–частоты
6	9	4–6	соблазны–уговоры–обещания–декларации; операционные системы–утилиты–драйверы–загрузка–выгрузка–погрузка;	обещания–уговоры; выгрузка–загрузка; драйверы–загрузка; утилиты–драйверы;	<i>обещания; уговоры; выгрузка; драйверы; загрузка;</i>	(1)уговоры (2)компьютеры (3)физика (4)сады–парки (5)лингвистика (6)изменения (7)медучреждения (8)охота (9)оружие
5	22	3–5	беспризорники–сироты–вдовы–вдовцы;	беспризорники–сироты; вдовы–вдовцы;	вдовы; сироты;	<фрагменты>
4	45	3–4	отличие–сходство–различие–повторение;	отличие–сходство; различие–повторение;	различие; сходство;	<фрагменты>
3	117	3	Тангаж–крен–дифферент;	тангаж–крен; крен–дифферент;	крен;	<фрагменты>
2	654	2	пассив–актив1;	пассив–актив1;	не приложимо	<фрагменты>

4. Заключение

Из баз пользовательских запросов поисковиков Google и Яндекс извлечена обширная совокупность сочиненных именных пар и на этой основе построена и описана ассоциативная сеть из 9,2 тыс. понятий, которые обычно входят парами в запросы к Интернету. У каждого из этих понятий в среднем по 2,7 ассоциативных партнеров. Тем самым предложена новая семантическая подсистема, базирующаяся на тех смысловых ассоциациях между понятиями, которые в большинстве не сводимы к связям типа WordNet.

Показано, что выявленные именные пары обычно входят с достаточной частотой и в текстовые массивы Интернета и потому, за редкими исключениями, могут считаться принадлежащими русскому языку в целом.

Созданную ассоциативную сеть удобно использовать при оперативном построении запросов к тем

поисковикам, которые либо пока не содержат данные пары понятий в своей базе пользовательских запросов, либо вообще не предоставляют пользователю доступ к этой базе. Можно также использовать эту сеть в программах классификации и установления семантической связности текстов.

Полезно было бы сравнить выявленные ассоциации с теми, которые известны по русскому ассоциативному словарю [8] и по идеографическому словарю типа тезауруса Роже [9]. Нечто интересное может выявиться при применении к полученной сети известных методов кластеризации. Когда станет известной аналогичная ассоциативная сеть для английского языка, можно будет сравнить интересы русскоязычных и англоязычных пользователей. Содержимое предложенной сети выложено на сайте [10].

Настоящая работа выполнена при частичной поддержке второго автора правительством Мексики (гранты SNI, SIP-IPN 20100773 и CONACYT 50206-H.)

Литература

1. *Большаков И. А.* КроссЛексика — большой электронный словарь сочетаний и смысловых связей русских слов. // Компьютерная лингвистика и интеллектуальные технологии. Международная конференция «Диалог 2009». Вып. 8 (15) М.: РГГУ, 2009. С. 45–50.
2. *Bolshakov I. A.* Stable Coordinated Pairs as a Specific Resource of Language. // East West Encounter: Second Intern. Conf. Meaning-Text Theory MTT'2005. Slavic Culture Languages Publ., Moscow, 2005, p. 86–97.
3. *Fellbaum Ch. (Ed.)* WordNet: An Electronic Lexical Database. // MIT Press, 1998.
4. *Leman F.* Semantic Networks. // Computer Math. Applic. Vol. 23, No. 2–5, 1992, p. 1–50.
5. *Kamps J., Marx M., Mokken R. J., de Rijke M.* Using WordNet to Measure Semantic Orientations of Adjectives. // Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC2004, vol. IV, 2004, p. 1115–1118.
6. *Erdos P., Renyi A.* On the evolution of random graphs. // Magyar Tudomanyos Akademia Matematikai Kutato Intezetenk Kozlomenyei. No.5, 1960, p. 17–61.
7. *Janson S., Luczak T., Rucinski A.* Random Graphs. // NY: Wiley — Interscience Series in Discrete Mathematics, 2000.
8. *Караулов Ю. Н., Сорокин Ю. А., Тарасова Е. Ф., Уфимцева Н. В., Черкасова Г. А.* Русский ассоциативный словарь. // М.: АСТ, 1994.
9. *Баранов О. С.* Идеографический словарь русского языка. // М.: ЭТС, 1995.
10. *Ассоциативная сеть русскоязычных пользователей Интернета.* www.Gelbukh.com/coordinate-pairs-in-queries/