Выявление синтагматических и парадигматических связей в морфологически размеченном корпусе текстов ♣

В.П. Захаров

Санкт-Петербургский государственный университет
vz1311@yandex.ru

Аннотация

Статья посвящена моделированию семантических отношений между лексическими единицами языка в целях формирования лексико-семантических микрополей как основы тезауруса (лексико-семантической системы) предметной области. Обсуждаются результаты экспериментов с использованием статистических методов на основе созданного корпуса специальных текстов.

В рамках работы над проектом РГНФ 10-04-00204а «Идея империи в русской литературе» (руководитель В.Е. Багно) была подготовлена библиотека (корпус) электронных текстов по теме. В состав библиотеки входят тексты XVIII — начала XX вв. общим количеством около 400 текстов. На основе библиотеки был создан корпус текстов объемом более 2 млн. слов, состоящий из четырех подкорпусов (1721–1800 гг., 1801–1856 гг., 1857–1905 гг., 1906–1924 гг.). Граничные даты подкорпусов выбраны как своего рода «вехи» в осознании понятия империи в развитии русской общественной мысли. Созданы словники подкорпусов и словник всего корпуса текстов и проведено исследование по выявлению характеристической лексики по теме и вычленению парадигматических отношений между лексемами (понятиями) с использованием автоматизированных методов.

Ставится задача моделирования семантики предметной (тематической) области знания (конкретно по теме указанного проекта) в терминах тезаурусного подхода и теории лексико-семантических полей. Используются статистические методы, в том числе дистрибутивный метод, и специальные программно-лингвистические инструменты.

Ряд специалистов [3, 4, 6] предлагают рассматривать тезаурус как модель терминологической системы. «Термин — слово или словесный комплекс, соотносящийся с понятием определенной организованной области познаний (науки, техники), вступающий в системные отношения с другими словами и словесными комплексами и образующее вместе с ними в любом отдельном случае и в определенное время замкнутую систему, отличающуюся высокой информативностью, однозначностью, точностью и экспрессивной нейтральностью» [6].

Терминосистему можно определить как, сложную динамическую и одновременно устойчивую систему, «элементами которой являются отобранные по определенным правилам лексические единицы какого-нибудь естественного языка, структура которой изоморфна структуре логических связей между понятиями специальной области знаний и деятельности, а функция состоит в том, чтобы служить знаковой (языковой) моделью этой области знаний и деятельности» [7].

Говоря в терминах лингвистики, речь идет о совокупности лексических единиц (имен понятий), которая обозначается терминами «лексическое поле», «лексико-семантическое поле», «функционально-семантическое поле» и т.д. Полем называют некоторое множество языковых единиц, объединенных по какому-то признаку или совокупности признаков. Приведем определение поля из словаря О.С. Ахмановой: «Поле — совокупность содержательных единиц, покрывающая определенную область человеческого опыта и образующая более или менее автономную микросистему» [2].

В трактовке В.Г. Адмони поле характеризуется наличием инвентаря элементов, связанных системными отношениями. В.Г. Адмони усматривает в поле центральную часть — ядро, элементы которого обладают полным набором признаков, определяющих данную группировку, и периферию, элементы которой обладают не всеми характерными для поля признаками, но могут иметь и признаки, присущие соседним полям [1]. Поле предполагает непрерывность связей объектов множества, причем на некоторых участках поля создаются области, в которых связи особенно интенсивны, а признаки особенно сильно выражены. При этом для поля характерна нечеткость границ между частями речи. В теории баз знаний терминосистема часто трактуется как онтология.

Важным аспектом является возможность введения метаопераций над лексикой предметной или тематической области. К такого рода метаопераци-

Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012), Санкт-Петербург, Россия, 2012.

ям относятся вычисления всевозможных метрик, а также более «интеллектуальные» семантические процедуры, базирующиеся на метриках и методологии дистрибутивно-статистического анализа. Как раз такой подход демонстрируется на материале данного исследования, а именно, ориентация на корпусы текстов и выявление семантических отношений на основе статистических метрик с последующей экспертной оценкой.

Специальные тексты во многом терминологичны, и поэтому должны быть разработаны принципы и методы автоматизированного выделения терминов и терминологических сочетаний (коллокаций), основанные на корпусных данных [5].

Большая часть лексических единиц терминосистемы или лексико-семантической системы предметной области представляет собой словосочетания. Поэтому встает задача выработки автоматических (полуавтоматических) методов выявления таких сочетаний по корпусам текстов. Многие из этих методов базируются на вероятностно-статистических основаниях. Статистический аппарат, применяемый в системах работы с корпусами текстов, позволяет пользователям ранжировать результаты обработки текстов по разным параметрам и задавать численные пороговые значения, что обеспечивает достоверность получаемых данных и создает параметрически настраиваемую систему.

За последние годы появилось большое число исследований и разработок, посвященных коллокациям (устойчивым сочетаниям), затрагивающих как теоретические аспекты статистического подхода к данному понятию, так и практические методы выявления коллокаций. Эти методы основываются на нахождении n-грамм (часто это биграммы или триграммы) в пределах заданного контекстного окна (диапазона). Самым простым способом выявления сочетаемости лексических единиц является составление частотных списков словосочетаний [9], т.е. частотных списков слов, оказавшихся слева или справа от ключевого слова. Математическим аппаратом для установления синтагматической связи между словами в тексте служат меры ассоциации (англ. association measures). В данном случае речь идет о статистической ассоциации, которая, в свою очередь, может иметь причиной синтаксическую или лексическую связанность. Линейная близость и частота совместной встречаемости могут оказаться важной предпосылкой для нахождения устойчивых сочетаний. Меры ассоциации вычисляют силу синтагматической связи между элементами в составе коллокации. Эти меры учитывают как частоту совместной встречаемости, так и другие параметры, прежде всего частоту в данном корпусе каждого отдельного элемента. Значения мер ассоциации можно считать показателями силы синтагматической связи между элементами словосочетаний.

При автоматическом выявлении словосочетаний сочетаний помимо статистических критериев отбора данных должны работать и другие методы, основывающиеся на собственно лингвистических моделях.

Существует подход к исследованию явления синтагматической связанности, который предполагает описание сочетаемости с помощью лексико-синтаксических шаблонов (иногда их называют лексико-грамматическими или морфологическими шаблонами) [8]. В данной статье нами будет использован метод выявления устойчивых сочетаний с использованием грамматики лексико-синтаксических шаблонов для русского языка. В нашем понимании, вслед за [8], лексико-синтаксический шаблон — это модель (структурный образец) языковой конструкции, в котором указываются существенные грамматические характеристики множества лексем, которые входят в языковые выражения, принадлежащие данному классу, и синтаксические условия употребления языкового выражения, построенного в соответствии с шаблоном (например, правила согласования морфологических признаков лексем).

Для работы с корпусами требуется особый программно-лингвистический инструментарий. В данной работе использована система Sketch Engine [10], в которой меры ассоциации вычисляются в рамках заданных лексико-синтаксических шаблонов. Нами была разработана грамматика словосочетаний для русского языка, которая встраивается в систему Sketch Engine. В соответствии с шаблонами вышеуказанной грамматики и на основе морфологически размеченного корпуса для заданных лексических единиц генерируются списки наиболее частотных (устойчивых) словосочетаний (лексические шаблоны, word sketches). Эти шаблоны отражают лексическую и синтаксическую сочетаемость этих единиц с другими словами с количественным указанием силы связи, которая рассчитывается на основе мер ассоциации [10]

В настоящий момент грамматика описывает следующие отношения:

- сочинительное отношение (***=и/или***);
- субъектное отношение (N_1+V: ***=subject/subject_of, =passive/subj_passive, =быть_adj/subj_быть***);
- объектное отношение (V+N_2, V+N_3, V+N_4, V+N_5: ***=object2/object2_of, =object3/object3_of, =object4/object4_of, =inst_modifier/inst_modifies;*** V+Vinf: ***=post_inf/verb_post_inf;*** $Adj_{кр}$+V: ***=modal_inf/modal***);
- атрибутивное отношение (N+N_2: ***=gen_modifier/gen_modifie;*** Adj+N ***=a_modifier/modifies***);
- компаративное отношение (N+Adj_{comp}+N_2: ***=comparative***);
- обстоятельственное отношение (***=adv_modifier/adv_modifies***);
- сочетания с предлогами (Prep+N, V+Prep: ***=prec_prep, =post_prep;*** N+PP, V+PP: ***=pp_%s, =pp_obj_%s***).

В системе Sketch Engine имеются специальные инструменты, позволяющие измерять силу как синтагматических, так и парадигматических связей на основе дистрибуции лексем в корпусе: *Лексические шаблоны*, *Тезаурус*, *Кластеризация* и *Дифференциация*. Указанные инструменты позволяют выявить термины, являющиеся словосочетаниями (синтагматика), и термины, образующие лексико-семантические поля (парадигматика).

По текстам указанного корпуса были созданы частотные словники, позволяющие выделить характеристическую лексику по теме. Приведем список тридцати наиболее частотных однословных терминов, встретившихся в корпусе: *Россия, народ, человек, церковь, жизнь, слово, время, мир, история, сила, государство, власть, век, год, начало, бог, дело, вера, земля, смысл, империя, мысль, имя, идея, Европа, дух, право, племя, отношение, страна*. Задача заключалась не просто в выявлении высокочастотных слов, а в определении лексического ядра ключевых слов по данной тематике, поэтому принимались во внимание как абсолютная частота терминов в корпусе, так и число документов, в которых встречается то или иное слово.

Был проведен анализ сочетаемости указанных ключевых слов в соответствии со следующей методикой. Для десяти из наиболее частотных существительных, включенных в базовый словарь, с помощью инструмента Тезаурус были вычислены по 100 слов, наиболее тесно с ним связанных. Далее был проведен анализ полученных 1000 слов. Данная совокупность насчитывает меньше, чем 1000 слов, так как в семантическом микрополе каждого из 10 исходных базовых терминов могут встречаться одни и те же единицы. Далее для каждого такого связанного слова из полученного множества подсчитывается частота вхождения в эти 10 семантических микрополей. Естественно предположить, что чем больше эта частота, тем в большей степени исследуемое слово является характерным для всего корпуса и тем самым для данной предметной области. Так, слова *«власть», «государство», «дух», Европа», «закон», «идея», «империя», «истина», «религия», «Рим»* присутствуют во всех 10 микрополях и тем самым входят в центральное ядро общего семантического поля по теме «Идея империи в русской культуре». Слова *«грех», «политика», «право», «просвещение», «путь», «роман», «сознание», «Христос»* встречаются в 6-7 микрополях и таким образом находятся на некотором семантическом расстоянии от ядра, а слова *«влияние», «война», «воля», «князь», «писатель», «раскол»* и др., представленные в 3-4 микрополях, должны быть отнесены к периферии данной терминосистемы.

Одновременно при создании своего рода «карты семантического поля» для каждого из этих анализируемых слов учитывается абсолютная частота вхождения в корпус и статистическая мера ассоциации, выданная инструментами «Тезаурус» и «Кластеризация» (см. ниже).

Подсистема (инструмент) Sketch Engine «Лексические шаблоны» (Word Sketches) позволяет выявлять устойчивые сочетания. Были проведены эксперименты с базовыми терминами. Ниже на рис. 1 приведен пример выдачи *словосочетаний (лексических шаблонов)* для ключевого слова «империя», отсортированных по синтаксическим моделям разработанной нами грамматики (элементы словосочетаний приводятся в нормализованной словарной форме).

В первом столбце каждой таблицы приведены слова, встречающиеся в контексте с ключевым словом «империя». Во втором столбце указана абсолютная частота того или иного словосочетания. В третьем столбце представлено значение статистической меры salience (подсчеты основаны на данных о частотах компонентов коллокаций), согласно которой выданы эти коллокации. Наиболее терминологичные атрибутивные словосочетания представлены в таблице «a_modifier»: *«Римская империя», «Российская империя», «Византийская империя», «Священная империя», «Православная империя», «Западная империя», «Восточная империя», «Османская империя», «мировая империя»* и др. Также для каждого выделенного словосочетания можно перейти к просмотру его контекстов.

Тезаурус в системе Sketch Engine (или, как его можно охарактеризовать, дистрибутивный тезаурус) позволяет увидеть, какие слова имеют схожую дистрибуцию с заданным словом. Для вычисления подобия слов рассматриваются наборы выданных списков сочетаемости для пар слов. Схожесть дистрибуции слов высчитывается статистически. Этот механизм в сочетании с инструментом *Кластеризация* позволяет строить кластеры лексических единиц, которые соответствуют лексико-семантическим группам.

империя *(noun)* **Rise_new_2 freq = 989** (1010.5 per million)

a_modifier	473	4.3
римский	117	12.03
российский	58	11.15
византийский	32	10.67
священный	23	10.33
православный	25	9.73
западный	20	9.4
восточный	14	9.39
османский	7	8.9
мировой	9	8.74
христианский	12	8.72
великий	12	8.49
советский	6	8.34
вселенский	6	8.29
бывший	5	8.22
германский	5	8.09
единый	5	8.09
многоконфессиональной	3	7.69
монгольский	3	7.61
универсальный	3	7.59
законный	3	7.54
новый	7	7.52
всемирный	3	7.41
последний	3	7.16
многонациональный	2	7.1
британский	2	7.08
петровский	2	7.05
русский	9	7.04

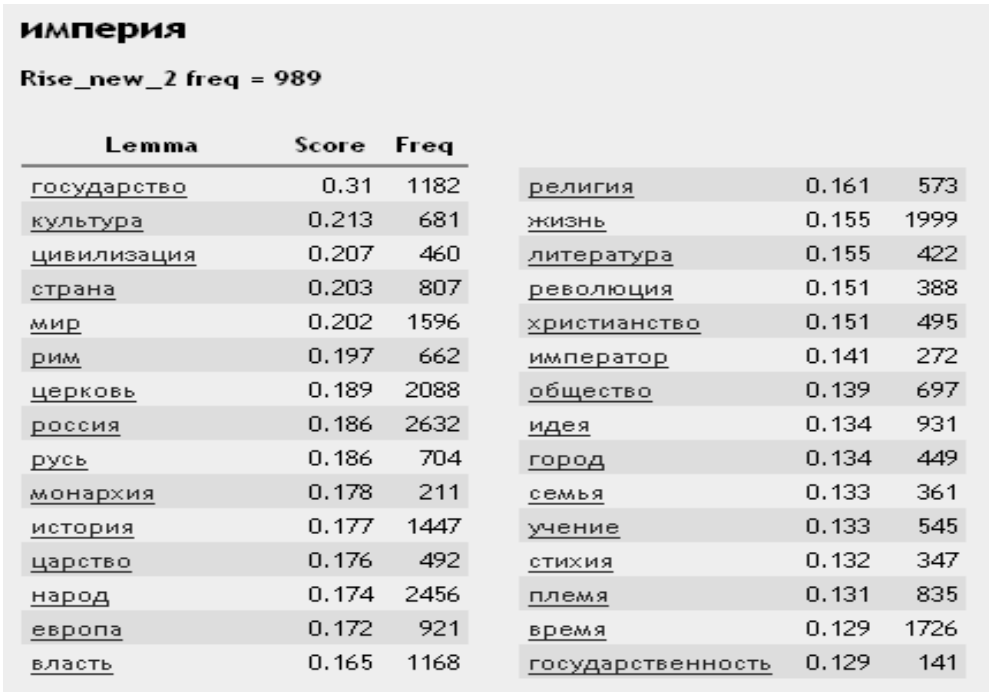
gen_modifies	427	3.0
распад	16	10.05
создание	13	9.5
принцип	13	9.15
идея	19	8.91
певец	6	8.78
перенос	6	8.77
существование	9	8.74
граница	8	8.61
территория	6	8.57
часть	13	8.55
идеолог	5	8.49
столица	6	8.38
гибель	5	8.24
разделение	5	8.24
император	6	8.21
вариант	4	8.1
становление	4	8.05
центр	5	8.03
основа	6	7.96
падение	5	7.96
онтология	3	7.81
христианизация	3	7.76
окраина	3	7.74
предел	4	7.74
область	5	7.71
крушение	3	7.68
модель	3	7.63

object4_of	36	2.4
восстановить	2	10.48
наводнять	1	9.79
сдать	1	9.75
возвести	1	9.71
разваливать	1	9.71
подорвать	1	9.71
поздравлять	1	9.71
добивать	1	9.68
возродить	1	9.64
взрывать	1	9.61
собрать	1	9.61
отличать	2	9.59
построить	2	9.59
освящать	1	9.39
потрясать	1	9.36
разрушить	1	9.33
населять	1	9.3
пережить	1	9.14
завоевывать	1	9.12
восстанавливать	1	9.12
строить	1	8.77
полагать	2	8.67
сравнивать	1	8.62
создавать	2	8.6
разделять	1	8.59
держать	1	8.54
создать	1	8.17

gen_modifier	67	0.5
Чингисхана	5	11.02
карл	5	10.54
ромеев	3	10.44
Габсбургов	2	9.85
нация	7	9.63
александр	4	9.62
романов	2	9.31
большевик	3	9.09
франк	2	8.93
Каролингов	1	8.91
Чингис-хана	1	8.89
Сассанидов	1	8.83
Чингизхана	1	8.83
Сасанидов	1	8.83
этносов	1	8.79
македонский	1	8.77
преступник	1	8.59
восток	4	8.46
зло	3	8.33
монгол	1	8.26
константин	1	7.9
антихрист	1	7.63
англия	1	7.12
запад	2	6.66
москва	1	6.4
царь	1	6.04
эпоха	1	5.83

Рис. 1. Пример выдачи словосочетаний для ключевого слова *«империя»*

На рис. 2 представлен результат выдачи для слова *«империя»*. В начале таблицы приводится заглавное (исследуемое) слово с его частотой в корпусе. Данные представлены в виде таблицы из трех столбцов: слово, значение статистической меры, частота слова в корпусе. Слова при этом упорядочены по значению статистической меры. Например, для лексемы *«мир»* значения равны 0,202 и 1596 соответственно.



империя

Rise_new_2 freq = 989

Lemma	Score	Freq
государство	0.31	1182
культура	0.213	681
цивилизация	0.207	460
страна	0.203	807
мир	0.202	1596
рим	0.197	662
церковь	0.189	2088
россия	0.186	2632
русь	0.186	704
монархия	0.178	211
история	0.177	1447
царство	0.176	492
народ	0.174	2456
европа	0.172	921
власть	0.165	1168
религия	0.161	573
жизнь	0.155	1999
литература	0.155	422
революция	0.151	388
христианство	0.151	495
император	0.141	272
общество	0.139	697
идея	0.134	931
город	0.134	449
семья	0.133	361
учение	0.133	545
стихия	0.132	347
племя	0.131	835
время	0.129	1726
государственность	0.129	141

Рис. 2. Пример выдачи механизма тезауруса для ключевого слова *«империя»*

Можно сказать, что список на рис. 2 содержит слова, входящие в одно лексико-семантическое поле со словом *«империя»*. Далее задача эксперта заключается в отборе слов, требующих более детального

описания, в том числе описания отношений, связывающих эти слова с заглавным словом и между собой. Как видим, в состав микрополя для термина *«империя»* вошли существительные, согласно статистической мере похожие по дистрибуции с данной лексемой (входят в одинаковые синтаксические отношения) и связанные с ней семантически: *«государство»*, *«культура»*, *«цивилизация»*, *«страна»*, *«мир»*, *«Рим»*, *«церковь»*, *«Россия»*, *«Русь»*, *«монархия»*, *«история»*, *«царство»* и др.

Покажем аналогичный результат для лексемы *«человек»* (рис. 3).

человек

Rise_new_2 freq = 2388

Lemma	Score	Freq			
народ	0.297	2456	европа	0.139	921
церковь	0.212	2088	власть	0.137	1168
государство	0.208	1182	вера	0.137	1025
мир	0.188	1596	христианин	0.136	282
россия	0.183	2632	дело	0.134	1104
бог	0.182	1138	русь	0.125	704
царь	0.177	525	история	0.124	1447
общество	0.168	697	религия	0.124	573
племя	0.155	835	христианство	0.123	495
жизнь	0.149	1999	литература	0.122	422
писатель	0.148	438	человечество	0.12	427
страна	0.147	807	душа	0.118	479
сила	0.147	1398	империя	0.117	989
личность	0.144	455	закон	0.117	723
дух	0.143	909	ум	0.115	341

Рис. 3. Пример выдачи механизма тезауруса для ключевого слова *«человек»*

Лексико-семантические поля для заданных терминов с разбиением на кластеры (микрополя) автоматически строит реализованная в системе функция *кластеризации*. Ниже приведен результат автоматического разбиения на кластеры лексем, связанных с лексемой *«империя»* (рис. 3).

империя

Rise_new_2 freq = 989

Lemma	Score	Freq	Cluster
государство	0.31	1182	страна [0.203, 807] европа [0.172, 921] религия [0.161, 573] общество [0.139, 697] человечество [0.121, 427]
культура	0.213	681	цивилизация [0.207, 460] литература [0.155, 422] просвещение [0.119, 317] философия [0.114, 254] наука [0.11, 395]
мир	0.202	1596	церковь [0.189, 2088] народ [0.174, 2456] племя [0.131, 835] человек [0.117, 2388]
рим	0.197	662	русь [0.186, 704] византия [0.122, 323]
россия	0.186	2632	
монархия	0.178	211	христианство [0.151, 495] православие [0.108, 435]
история	0.177	1447	жизнь [0.155, 1999] развитие [0.127, 805]
царство	0.176	492	
власть	0.165	1168	
революция	0.151	388	война [0.12, 449]
император	0.141	272	царь [0.124, 525]
идея	0.134	931	политика [0.114, 238] мысль [0.109, 961]
город	0.134	449	раскол [0.116, 172]
семья	0.133	361	нация [0.128, 238]
учение	0.133	545	дух [0.12, 909] вера [0.116, 1025]

Рис. 4. Гнездо тезауруса с выделенными кластерами для ключевого слова *«империя»*

В первом столбце приведены лексемы, во втором — значение статистической меры logDice, в третьем — абсолютная частота лексемы в корпусе, в четвертом — лексемы, образующие с лексемой из первого столбца единый кластер (в квадратных скобках указаны значение меры и частота, им соответствующие). При автоматической кластеризации были выделены следующие группировки слов (состоят из двух или более элементов): 1) «*государство*», «*страна*», «*Европа*», «*религия*», «*общество*», «*человечество*»; 2) «*культура*», «*цивилизация*», «*литература*», «*просвещение*», «*философия*», «*наука*»; 3) «*мир*», «*церковь*», «*народ*», «*племя*», «*человек*»; 4) «*Рим*», «*Русь*», «*Византия*»; 5) «*Россия*»; 6) «*монархия*», «*христианство*», «*православие*»; 7) «*история*», «*жизнь*», «*развитие*»; 8) «*царство*»; 9) «*власть*»; 10) «*революция*», «*война*»; 11) «*император*», «*царь*»; 12) «*идея*», «*политика*», «*мысль*»; 13) «*город*», «*раскол*»; 14) «*семья*», «*нация*»; 15) «*учение*», «*дух*», «*вера*». Можно заметить, что внутри кластеров сила и природа парадигматических связей разная: встречаются слова-синонимы, которые взаимозаменяемы в ряде контекстов, и слова, связанные другими отношениями. Примером первого типа могут служить лексемы «*император*» и «*царь*», «*государство*» и «*страна*».

На рис. 5 приведены результаты кластеризации для лексемы «*человек*».

человек

Rise_new_2 freq = 2388

Lemma	Score	Freq	Cluster
народ	0.297	2456	церковь [0.212, 2088] бог [0.182, 1138] племя [0.155, 835]
государство	0.208	1182	мир [0.188, 1596] страна [0.147, 807] европа [0.139, 921] религия [0.124, 573] империя [0.117, 989]
россия	0.183	2632	
царь	0.177	525	христос [0.114, 486]
общество	0.168	697	человечество [0.12, 427] природа [0.089, 320]
жизнь	0.149	1999	история [0.124, 1447] литература [0.122, 422] культура [0.111, 681] цивилизация [0.11, 460]
писатель	0.148	438	мыслитель [0.097, 196] поэт [0.097, 260]
сила	0.147	1398	дух [0.143, 909] свобода [0.114, 779] истина [0.111, 674] начало [0.11, 1140]
личность	0.144	455	
власть	0.137	1168	
вера	0.137	1025	любовь [0.099, 480]
христианин	0.136	282	
дело	0.134	1104	закон [0.117, 723] идея [0.093, 931]
русь	0.125	704	рим [0.107, 662] византия [0.092, 323]
христианство	0.123	495	православие [0.094, 435]

Рис. 5. Гнездо тезауруса с выделенными кластерами для ключевого слова «*человек*»

Таким образом, мы видим, что использование корпуса текстов и инструментов системы Sketch Engine позволяет выявлять в автоматизированном режиме синтагматические и парадигматические связи и создавать более адекватное наполнение терминосистемы.

Как методы выявления коллокаций на основе статистических мер ассоциации, так и инструмент «Лексические шаблоны» в составе Sketch Engine (выявление коллокаций, распределенных по синтаксическим моделям) являются мощным средством выявления устойчивых сочетаний разного типа. Естественно, последнее слово остается за экспертом, решающим вопрос, какие словосочетания являются понятиями и базовыми элементами терминосистемы.

Другая особенность системы Sketch Engine заключается в том, что в ней, как уже говорилось, имеются средства, реализующие методику дистрибутивно-статистического анализа — «Тезаурус», «Кластеризация» и «Дифференциация». Все они, разными способами, выявляют парадигматические (т.е. семантические) связи между терминами с количественным указанием силы этой связи. После этого задача эксперта — специалиста в данной предметной области — определить и явно задать, если требуется, тип этих связей.

Эксперименты по выявлению отношений между словами на базе текстового корпуса, в свою очередь, позволяют наметить и пути совершенствования системы Sketch Engine. Одним из таких направлений может стать написание грамматики для указанной системы для работы с семантически размеченным корпусом.

Литература

[1] Адмони В. Г. Синтаксис современного немецкого языка: Система отношений и система построения. — Л.: Наука, 1973. — 366 с.
[2] Ахманова О.С. Словарь лингвистических терминов. — М.: Советская энциклопедия, 1966. — 608 с.
[3] Головин Б.Н. Типы терминосистем и основания их различения // Термин и слово. — Горький: ГГУ, 1981. — С. 3–8.
[4] Головин Б.Н., Кобрин Р.Ю. Лингвистические основы учения о терминах. — М.: Высш. шк., 1987. — 104 с.
[5] Захаров В.П. Тезаурус по корпусной лингвистике // Информационные технологии и письменное наследие. El'Manuscript-10. Материалы Международной научной конференции. — Уфа, 2010. — С. 95-98.
[6] Кияк Т.Р. Лингвистические аспекты терминоведения: Учебное пособие. — Киев: УМКВО, 1989. — 103 с.
[7] Лейчик В.М., Смирнов И.П., Суслова И.М. Терминология информатики: теоретические и практические вопросы // Итоги науки и техники. — М.: ВИНИТИ, 1977. — Т. 2. — С. 40-53.
[8] Митрофанова О.А., Захаров В.П. Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). — М.: РГГУ, 2009. — С. 321—328.
[9] Пиотровский Р.Г. Текст, машина, человек. — Л.: Наука, 1975. — 327 с.
[10] Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress. — Lorient: Universite de Bretagne-Sud, 2004. — P. 105-116.

Studying Syntagmatic and Paradigmatic Relations in a Morphologically Annotated Corpus

V. Zakharov

The paper deals with modeling semantic relations between lexical items. The paper presents the results of automatic term extraction from a special text corpus. The method applied includes statistical analysis that enables estimating paradigmatic and syntagmatic relations between lexemes based on their distribution.

♣ Исследование выполнено при финансовой поддержке РГНФ в рамках научно-исследовательского проекта РГНФ («Идея Империи в русской литературе»), проект № 10-04-00204а.