ЛЕКСИКО-СИНТАКСИЧЕСКИЕ ШАБЛОНЫ КАК ИНСТРУМЕНТ ВЫЯВЛЕНИЯ СПЕЦИАЛЬНОЙ ЛЕКСИКИ ПРЕДМЕТНОЙ ОБЛАСТИ

LEXICO-SYNTACTIC PATTERNS AS A TOOL FOR EXTRACTING LEXIS OF A SPECIALIZED KNOWLEDGE DOMAIN

Хохлова М.В. (*khokhlova.marie@gmail.com*)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

В статье обсуждаются результаты экспериментов по выявлению терминов на базе корпуса специальных текстов. Рассматривается подход к исследованию явления синтагматической связанности, который предполагает описание сочетаемости с помощью лексико-синтаксических шаблонов в рамках системы Sketch Engine. В ходе работы проведен анализ лексического состава текстов по корпусной лингвистике с использованием статистических методов, позволяющих выявить парадигматические и синтагматические отношения на основе дистрибуции лексем.

Ключевые слова: специальный текст, корпус, дистрибутивно-статистические методы, коллокации, русский язык, автоматическое извлечение терминов, тезаурус

Khokhlova M.V. (*khokhlova.marie@gmail.com*)

St.Petersburg State University, St.Petersburg, Russia

The paper presents the results of automatic term extraction from a specialized text corpus (a collection of papers on corpus linguistics) by means of statistical methods (association measures) combined with certain syntactic models. The approach undertaken in the paper is based on lexico-syntactic patterns that can be viewed as models of phrases for the Russian language. Sketch Engine represents a corpus tool which takes as input a corpus of any language and corresponding grammar patterns. The system gives information about a word's collocability on concrete dependency models, and generates lists of the most frequent phrases for a given word, based on appropriate models. During our work we have written grammatical rules that take into account syntactic constructions of the Russian language based on the morphologically tagged corpus. The extracted terms belong to various clusters and represent the lexical structure of the texts in question. The applied method includes statistical analysis that enables estimating paradigmatic and syntagmatic relations between lexemes based on their distribution.

Keywords: specialized text, corpora, distributional and statistical methods, collocations, the Russian language, automatic term extraction, thesaurus

Целью настоящего исследования является моделирование парадигматических и синтагматических отношений на материале корпуса специальных текстов русского языка в целях формирования лексико-семантических микрополей как основы тезауруса предметной области.

Большая часть терминов — лексических единиц терминосистемы и тезауруса 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.

На вероятностный характер языка указывают многие исследователи. В корпусной лингвистике принято опираться на совместную встречаемость языковых единиц. За последние годы появилось большое число исследований и разработок, посвященных коллокациям, затрагивающих как теоретические аспекты статистического подхода к данному понятию, так и практические методы выявления коллокаций (см., например, обзор в работе (Evert 2004)). В ряде случаев для работы с корпусами специальных текстов требуется особый лингвистический инструментарий как на входе, так и на выходе корпуса. Это определяется как особенностями документов, так и задачами исследования. В данной статье нами будет использован метод выявления устойчивых терминологических сочетаний с использованием грамматики лексико-синтаксических шаблонов для русского языка. В нашем понимании, вслед за (Митрофанова, Захаров 2008), лексико-синтаксический шаблон — это структурный образец (модель) 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).

Грамматика лексико-синтаксических шаблонов встраивается в систему *Sketch Engine*, разработанную английскими и чешскими исследователями и оперирующую понятием «лексических портретов» (*word sketches*), фиксирующих лексическую и грамматическую сочетаемость лексических единиц (Kilgarriff et al. 2004; Хохлова 2010). На основе морфологически размеченного корпуса данная система порождает списки слов, в которых содержится информация об их «лингвистическом поведении» — сочетаемости с другими словами с количественным указанием силы связи, которая рассчитывается на основе известных мер ассоциации. Результат работы программы представлен наиболее частотными (устойчивыми) словосочетаниями, расклассифицированными по типам в соответствии с лексико-синтаксическими шаблонами вышеуказанной грамматики. Всего нами в рамках исследования было описано 18 типов отношений, среди них:

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_modifies;*** 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***).

Далее данная грамматика была встроена в систему Sketch Engine.

В системе Sketch Engine имеются специальные инструменты, позволяющие измерять силу не только синтагматических, но и парадигматических связей на основе дистрибуции лексем в корпусе: *тезаурус* (thesaurus), *кластеризация* (clustering) и *дифференциация* (differences), которые будут нами рассмотрены ниже. Работа данных инструментов основывается как на статистических критериях, так и разработанных нами для русского языка лексико-синтаксических шаблонах.

В ходе исследования был использован специальный корпус, который включает в себя русскоязычные тексты по корпусной лингвистике, собранные на кафедре математической лингвистики СПбГУ (руководитель проекта — В.П. Захаров). В него вошли материалы международных конференций «Корпусная лингвистика и лингвистические базы данных–2002» (Санкт-Петербург, 5–7 марта 2002 г.), «Корпусная лингвистика–2004» (Санкт-Петербург, 11–14 октября 2004 г.), «Корпусная лингвистика–2006» (Санкт-Петербург, 10–14 октября 2006 г.), «Корпусная лингвистика–2008» (Санкт-Петербург, 6–10 октября 2008 г.) и «Корпусная лингвистика–2011» (Санкт-Петербург, 26–29 июня 2011 г.), а также учебник по корпусной лингвистике (Захаров, Богданова 2011). В настоящее время корпус насчитывает более 300 тыс. словоупотреблений.

Приведем список тридцати наиболее частотных однословных терминов, встретившихся в корпусе: *текст, корпус, слово, язык, словарь, данные, система, значение, тип, разметка, анализ, предложение, форма, исследование, работа, время, единица, глагол, случай, часть, информация, создание, речь, материал, структура, существительное, база, пример, задача, связь*.

Наша задача — вычислить на основе дистрибуции силу синтагматических и парадигматических связей между словами базовой лексики, т.е. выявить термины, являющиеся словосочетаниями, и термины, входящие в одно лексико-семантическое поле. Были проведены эксперименты с частотными терминами указанной предметной области. Ниже приводятся некоторые результаты по терминологии корпусной лингвистики с использованием вышеназванных инструментов.

Тезаурус в системе Sketch Engine (или, как его можно охарактеризовать, дистрибутивный тезаурус) позволяет увидеть, какие слова встречаются с теми же словами, что и ключевое слово. Таким образом, пользователь получает данные о том, какие слова в корпусе имеют дистрибуцию, схожую с заданным словом. Для вычисления подобия слов рассматриваются наборы списков сочетаемости для слов X1 и X2. Схожесть дистрибуции слов высчитывается на основе статистической меры logDice (Kilgarriff et al. 2004). Неинформативные случаи, для которых значение меры отрицательно, отбрасываются. Рассматриваются словосочетания, в которых слова X1 и X2 встречаются в одинаковых грамматических контекстах (например, объектное отношение при одном глаголе). Этот механизм в результате позволяет формировать группы лексических единиц, которые соответствуют лексико-семантическим микрополям.

Результат выдачи для слова *«корпус»* (см. рис. 1) представлен в виде таблицы из трех столбцов: слово, значение статистической меры, частота слова в корпусе. Слова при этом упорядочены по значению статистической меры. В начале таблицы приводится заглавное (исследуемое) слово с его частотой.

корпус

Corpus Linguistics freq = 2708

Lemma	Score	Freq			
текст	0.298	3220	предложение	0.113	574
словарь	0.259	918	источник	0.112	199
язык	0.227	1782	коллекция	0.112	128
материал	0.209	452	контекст	0.112	347
система	0.195	768	единица	0.108	492
база	0.193	421	форма	0.106	566
слово	0.178	1814	документ	0.106	230
анализ	0.161	585	перевод	0.104	384
разметка	0.159	615	объем	0.101	268
часть	0.131	476	пример	0.099	402
структура	0.127	439	массив	0.098	115
тип	0.125	648	глагол	0.095	489
описание	0.124	290	результат	0.094	388
использование	0.122	371	модель	0.094	212
создание	0.117	461	метод	0.092	213
исследование	0.114	562			

Рис. 1. *Пример выдачи механизма «Тезаурус» для ключевого слова «корпус»*

Можно сказать, что список на рис. 1 содержит слова, входящие в одно лексико-семантическое поле со словом *«корпус»*. Далее задача эксперта, в данном случае, корпусного лингвиста, выявить и назвать типы тезаурусных (семантических) отношений, связывающих эти слова с заглавным словом и между собой.

Покажем состав полученного микрополя для термина *«разметка»* (рис. 2). Сюда вошли существительные, имеющие согласно статистической мере дистрибуцию, схожую с данной лексемой (входят с лексемой *«разметка»* в одинаковые синтаксические отношения): «*анализ», «обработка», «описание», «перевод», «создание», «исследование», «построение», «тип»* и др.

разметка

Corpus Linguistics freq = 615

Lemma	Score	Freq			
анализ	0.36	585	словарь	0.156	918
обработка	0.291	238	классификация	0.151	142
описание	0.241	290	выравнивание	0.148	69
перевод	0.221	384	объем	0.146	268
создание	0.22	461	определение	0.144	187
исследование	0.195	562	выделение	0.144	103
построение	0.183	131	разработка	0.143	160
тип	0.182	648	работа	0.141	517
представление	0.178	207	количество	0.141	248
использование	0.177	371	характеристика	0.137	284
структура	0.175	439	снятие	0.136	80
подготовка	0.165	66	язык	0.136	1782
поиск	0.159	368	выражение	0.136	140
корпус	0.159	2708	информация	0.135	467
база	0.157	421	предложение	0.133	574

Рис. 2. *Микрополе (тезаурусное гнездо) для ключевого слова «разметка»*

Кроме инструмента «Тезаурус», лексико-семантические микрополя для заданных терминов позволяет строить *функция кластеризации,* реализованная в системе. Ниже приведен результат автоматической разбиения на кластеры лексем, связанных с лексемой *«разметка»* (см. рис. 3).

разметка

Corpus Linguistics freq = 615

Lemma	Score	Freq	Cluster
анализ	0.36	585	обработка [0.291, 238] исследование [0.195, 562]
описание	0.241	290	представление [0.178, 207] классификация [0.151, 142]
перевод	0.221	384	выделение [0.144, 103]
создание	0.22	461	построение [0.183, 131] подготовка [0.165, 66] разработка [0.143, 160] формирование [0.109, 74]
тип	0.182	648	
использование	0.177	371	
структура	0.175	439	значение [0.112, 671]
поиск	0.159	368	
корпус	0.159	2708	язык [0.136, 1782] текст [0.117, 3220]
база	0.157	421	материал [0.105, 452]
словарь	0.156	918	
выравнивание	0.148	69	сравнение [0.124, 104]
объем	0.146	268	количество [0.141, 248] число [0.111, 384]
определение	0.144	187	выражение [0.136, 140]
работа	0.141	517	
характеристика	0.137	284	признак [0.117, 330]
снятие	0.136	80	извлечение [0.111, 46]
информация	0.135	467	

Рис. 3. *Гнездо тезауруса с выделенными кластерами для ключевого слова «разметка»*

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) *«анализ», «обработка», «исследование»; 2) «описание», «представление», «классификация»; 3) «перевод», «выделение»; 4) «создание», «построение», «подготовка», «разработка», «формирование»; 5) «структура», «значение»; «корпус», «язык», «текст»; 6) «база», «материал»; 7) «выравнивание», «сравнение», «объем», «количество», «число»; 8) «определение», «выражение», «характеристика», «признак»; 9) «снятие», «извлечение».* Можно заметить, что внутри кластеров сила и природа парадигматических связей разная: встречаются синонимы, которые взаимозаменяемы в ряде контекстов, и слова, связанные другими отношениями. Примером первого типа могут служить лексемы *«характеристика»* и *«признак»*.

На рис. 4 приведен результат кластеризации существительных, связанных с лексемой *«текст»*.

текст

Corpus Linguistics freq = 3220

Lemma	Score	Freq	Cluster
корпус	0.298	2708	словарь [0.181, 918] материал [0.169, 452] система [0.144, 768] база [0.104, 421]
язык	0.267	1782	
слово	0.215	1814	единица [0.122, 492] глагол [0.112, 489] лексема [0.077, 166] существительное [0.075, 432]
предложение	0.173	574	часть [0.142, 476]
документ	0.151	230	файл [0.068, 128]
контекст	0.137	347	характеристика [0.079, 284]
разметка	0.117	615	
пример	0.116	402	случай [0.085, 487]
информация	0.111	467	
значение	0.11	671	структура [0.108, 439] тип [0.105, 648]
речь	0.1	461	
конструкция	0.098	284	словосочетание [0.076, 208] термин [0.063, 200]
форма	0.093	566	
версия	0.09	94	лексика [0.072, 140]
источник	0.086	199	объект [0.076, 232]
ошибка	0.084	168	
отношение	0.083	354	связь [0.074, 392]

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

С уверенностью можно назвать следующие кластеры: 1) «источник исследования» — *«корпус», «словарь», «материал», «система», «база»*; 2) «объект исследования» — *«слово, «единица», «глагол», «лексема», «существительное»*; 3) «конструкции» — *«конструкция», «словосочетание», «термин»*; 4) «отношение» — *«отношение», «связь»*. Среди лексем, не вошедших в кластеры, выделяются следующие: *«язык», «разметка», «информация», «речь», «форма», «ошибка»*.

Рассмотрим также возможные сложные термины, компонентами которых являются наиболее частотные слова в корпусе. Ниже представлены типичные словосочетания для некоторых частотных лексем (отсортированы по уменьшению их частоты), выданные согласно лексико-синтаксическим шаблонам аппарата Word Sketch:

слово

1) ***Adj N*** — *зависимое слово, целевое слово, ключевое слово, знаменательное слово, фонетическое слово, неопознанное слово, служебное слово, реестровое слово, запрошенное слово, измененяемое слово, анализируемое слово, фонетическое слово, полурусское слово, интересуемое слово, заимствованное слово, сленговое слово, изменяемое слово, неизменяемое слово, полноударное слово, заглавное слово, нейтральное слово, последующее слово, главное слово, новое слово, простое слово, исходное слово, сложное слово, конкретное слово, русское слово* и др.;

2) ***N N_2*** — *значение слова, употребление слова, характеристика слова, форма слова, группа слов, поиск слова, класс слов, смысл слова, признак слова, список слов,*

семантика слова, порядок слова, количество слов, принадлежность слова, написание слов, связь слов, уровень слов, пара слов, категория слова, число слова, многозначность слова, частотность слова, валентность слова, сочетаемость слов и др.;

разметка

1) Adj N *— морфологическая разметка, семантическая разметка, синтаксическая разметка, автоматическая разметка, лингвистическая разметка, грамматическая разметка, структурная разметка, библиографическая разметка, полная разметка, экстралингвистическая разметка, метатекстовая разметка, просодическая разметка, стандартная разметка, автоматизированная разметка, морфосинтаксическая разметка, частеречная разметка, морфемная разметка, многоуровневая разметка, дискурсивная разметка, однозначная разметка, полуавтоматическая разметка, ручная разметка, предварительная разметка, тематическая разметка* и др.;

2) ***N N_2*** *— способ разметки, схема разметки, тип разметки, глубина разметки, этап разметки, технология разметки, техника разметки, пример разметки, инструмент разметки, просмотр разметки, проблема разметки, процедура разметки, принцип разметки, система разметки, программа разметки, вид разметки, процесс разметки, скорость разметки, понятие разметки, результат разметки, формат разметки, качество разметки, вариант разметки, уровень разметки* и др.;

глагол

1) ***Adj N*** *— фразовый глагол, нулевой глагол, модальный глагол, вспомогательный глагол, каузативный глагол, префиксальный глагол, стативный глагол, непереходный глагол, переходный глагол, многозначный глагол, значимый глагол, английский глагол, украинский глагол, русский глагол* и др.;

2) ***N N_2*** *— форма глагола, валентность глагола, употребление глагола, сосед глагола, время глагола, употребление глагола, противопоставление глагола, причастие глагола, преобразование глагола, значение глагола, подсчет глаголов, особенность глагола, класс глагола, пара глаголов, реализация глагола, сочетание глагола, свойство глагола, наличие глагола, группа глаголов, коллокат глагола, спряжение глагола, квазисинонимичность глагола, опущение глагола, синтагма глагола, квазиоснова глагола, подгруппа глагола, помета глагола, аргумент глагола, залог глагола, рамка глагола, лицо глагола* и др.

Отметим, что среди приведенных словосочетаний есть как термины, которые могут пополнить тезаурус или словарь (например, *знаменательное слово, заглавное слово, морфологическая разметка, фразовый глагол, модальный глагол* и др.), так и просто высокочастотные сочетания (например, *конкретное слово, многозначность слова,*

процедура разметки и др.). В последней группе возможно выделение сочетаний разной степени устойчивости (например, свободные — *английский глагол, украинский глагол, русский глагол*; постоянно воспроизводимые — *реализация глагола, опущение глагола*). И те, и другие словосочетания могут служить материалом при описании терминосистемы данной области науки.

Функция «Дифференциация» (рис. 5) позволяет визуально показывать результат сравнения контекстов для пары слов, похожих по своей дистрибуции. При этом показываются модели и словосочетания, присущие обоим словам, а также те модели и словосочетания, которые характерны (и в какой степени характерны) для каждого из данных слов. В крайнем случае, несмотря на сходство дистрибуций, какие-то модели или словосочетания для одного из исследуемых слов являются единственно возможными.

gen_modifies	215	213	1.4	1.4
ход	0	10	0.0	10.2
проведение	0	4	0.0	8.9
алгоритм	0	4	0.0	8.7
метод	1	17	6.2	10.3
результат	5	26	8.1	10.5
модуль	1	5	6.7	9.1
основа	3	11	7.4	9.3
программа	5	14	8.3	9.8
процедура	4	10	8.4	9.7
вариант	4	9	8.0	9.2
процесс	4	6	8.1	8.7
этап	5	4	8.7	8.4
инструмент	4	2	8.6	7.6
техника	3	1	8.6	7.0
пример	6	2	8.6	7.0
технология	4	1	8.7	6.7
тип	14	3	8.9	6.6
схема	5	1	9.0	6.7
глубина	3	0	8.7	0.0
способ	6	0	9.1	0.0

a_modifier	287	239	3.8	3.2
контекстный	0	8	0.0	9.9
количественный	0	8	0.0	9.8
корпусный	0	5	0.0	9.3
графематический	0	4	0.0	9.1
терминологический	0	4	0.0	8.9
качественный	0	4	0.0	8.9
дискриминантного	0	3	0.0	8.7
прецедентного	0	3	0.0	8.7
сравнительный	0	3	0.0	8.6
статистический	1	10	6.5	10.0
тематический	2	3	7.6	8.5
морфологический	64	47	11.7	11.3
синтаксический	35	19	11.0	10.2
семантический	48	17	11.2	9.8
автоматический	20	6	10.4	8.8
грамматический	11	2	9.6	7.3
лингвистический	16	2	9.8	6.9
полный	5	0	8.7	0.0
Библиографическая	4	0	8.8	0.0
структурный	5	0	9.0	0.0

Рис. 5. *Пример выдачи функции «Дифференциация» для ключевых слов «разметка» и «анализ»*

В рассматриваемых таблицах каждому слову соответствует четыре статистических характеристики: первые две — частота встречаемости словосочетаний для слов *«разметка»* и *«анализ»* (первая и вторая позиции соответственно); вторые два — значения меры ассоциации для каждого словосочетания. Лексемы, приведенные на светлом фоне,

— например, *«тематический», «морфологический», «синтаксический», «семантический», «автоматический»* — сочетаются в равной степени с лексемами *«разметка»* и *«анализ»*. Лексемы *«контекстный», «количественный», «корпусный»* и др. (на темном фоне) встречаются только с существительным *«анализ»* (количество сочетаний с лексемой *«разметка»* равно 0). Прилагательные *«грамматический»* и *«лингвистический»* встречаются преимущественно со словом *«разметка»* (11 и 16 контекстов против 2 и 2 контекстов соответственно), *«статистический»* — преимущественно со словом *«анализ»* (10 контекстов против 1) и т.д.

Таким образом, мы видим, что использование корпуса и инструментов системы Sketch Engine, обладающей рядом уникальных инструментов, позволяет выявлять в автоматизированном режиме синтагматические и парадигматические связи и создавать более адекватное наполнение тезауруса и терминосистемы. Следует отметить, что основная масса понятий той или иной предметной области выражается словосочетаниями. Это справедливо и для предметной области «корпусная лингвистика». Как методы выявления коллокаций на основе статистических мер ассоциации, так и инструмент Word Sketch (коллокации, распределенные по синтаксическим моделям) в составе Sketch Engine являются мощным средством выявления устойчивых сочетаний разного типа. Естественно, последнее слово остается за экспертом, какие словосочетания следует включать в тезаурус. 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. Опять же задача эксперта — специалиста в данной предметной области — определить и явно задать 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

Evert S. (2004), *The statistics of Word Cooccurences. Word Pairs and Collocations. PhD thesis*, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung (IMS), Stuttgart.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine. *Proceedings of the XIth Euralex International Congress.* Universite de Bretagne-Sud, Lorient, 2004, pp. 105-116.

Doklady nauchnoj konferentsii "Korpusnaja lingvistika i lingvisticheskie bazy dannyh" [Proceedings of the International Conference "Corpus Linguistics and Linguistic Databases"] (2002), Izd-vo St.Peterburgskogo universiteta, St.Petersburg.

Hohlova M.V. (2010), Writing the Grammatical Module for the Russian Language for a Specialized Corpus Query System [Razrabotka grammaticheskogo modulja russkogo jazyka dlja spetsializirovannoj sistemy obrabotki korpusnyh dannyh], *Vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta* [Herald of the St.Petersburg State University], Serija 9. Filologija, vostokovedenie, zhurnalistika. Izd-vo St.Peterburgskogo universiteta, St.Petersburg. Vol. 2, pp. 162-169.

Mitrofanova O.A., Zaharov V.P. Automatic Analysis of Terminology in a Russian Corpus of Texts on Corpus Linguistics [Avtomatizirovannyj Analiz Terminologii v Russkojazychnom Korpuse Tekstov po Korpusnoj Lingvistike]. *Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009"* [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009"]. Bekasovo, 2009, pp. 321—328.

Trudy Mezhdunarodnoj Konferentsii "Korpusnaja Lingvistika–2004" [Proceedings of the International Conference "Corpus Linguistics–2004"] (2004), Izd-vo St. Peterburgskogo universiteta, St.Petersburg.

Trudy Mezhdunarodnoj Konferentsii "Korpusnaja Lingvistika–2006" [Proceedings of the International Conference "Corpus Linguistics–2006"] (2006), Izd-vo St. Peterburgskogo universiteta, St.Petersburg.

Trudy Mezhdunarodnoj Konferentsii "Korpusnaja Lingvistika–2008" [Proceedings of the International Conference "Corpus Linguistics–2008"] (2008), Izd-vo St. Peterburgskogo universiteta, St.Petersburg.

Trudy Mezhdunarodnoj Konferentsii "Korpusnaja Lingvistika–2011" [Proceedings of the International Conference "Corpus Linguistics–2011"] (2011), Izd-vo St. Peterburgskogo universiteta, St.Petersburg.

Zaharov V.P. Thesaurus on Corpus Linguistics [Tezaurus po korpusnoj lingvistike]. *Informatsionnye tehnologii i pis'mennoe nasledie. El'Manuscript-10. Materialy mezhdunarodnoj nauchnoj konferentsii* [Information Technologies and Textual Heritage. El'Manuscript-10. Proceedings of the International Scientific Conference]. Ufa, 2010, pp. 95-98.

Zaharov V.P., Bogdanova S.Ju. (2011), *Korpusnaja lingvistika* [Corpus Linguistics]. IGLU, Irkutsk.