Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке[1]

Study of effectiveness of statistical measures for collocation extraction on Russian texts

Захаров В. П. (vz1311@yandex.ru),
Хохлова М. В. (vertikal-maria@yandex.ru)

Санкт-Петербургский государственный университет
Институт лингвистических исследований РАН, Санкт-Петербург

Аннотация. Описаны результаты исследования по выявлению устойчивых сочетаний в русском языке. Эксперимент заключался в нахождении и анализе биграмм с частотными глаголами и существительными русского языка. Цель — изучить сочетаемостные характеристики данных лексических единиц, соотнести результаты, полученные на основе различных мер ассоциации на разных корпусах, сравнить наиболее популярные меры ассоциации. Рассматриваются требования к программному обеспечению.

1. Понятие коллокации в лингвистике

Термин «коллокация», хотя и вошел в постоянное употребление сравнительно недавно, по праву занимает одно из ключевых мест в современной лингвистике. В широком смысле это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости. Коллокации в настоящий момент играют ведущую роль в лексикографической практике (Atkins 2008; Hausmann 1979; Hausmann 1985; Kilgarriff 2006; Sinclair 1991). В последнее время за рубежом и в России создаются специальные словари коллокаций (Benson 1986; Crowther et al. 2002; Kjellmer 1994; Krishnamurthy et al. 2006; Sinclair 1995; Бирюк 2008; Денисов 2002, Кустова и др. 2008).

Однако существующие словари устойчивых словосочетаний, во-первых, охватывают далеко не полный их перечень, во-вторых, часто делают это недостаточно последовательно. Особенно это справедливо для русского языка. Поэтому актуальность работ по автоматическому выявлению коллокаций из текстов несомненна.

В настоящее время мы видим несколько важнейших прикладных задач, где есть нужда в автоматизированных методах извлечения коллокаций из больших корпусов текстов. В частности, это составление словарей и других лексикографических пособий, составление онтологий, обучение языку, отладка лингвопроцессоров, задачи информационного поиска.

Кратко коснемся самого понятия коллокация. Существуют различные определения этого понятия. В целом в основе большинства определений коллокации лежит явление семантико-грамматической взаимообусловленности элементов словосочетания (см., напр., (Иорданская, Мельчук 2007)).

Термин «коллокация» в русскоязычной научной литературе впервые появился в Словаре лингвистических терминов О. С. Ахмановой (Ахманова 1966). Первой работой в российской лингвистике, полностью посвященной исследованию понятия коллокации на материале русского языка, является монография Е. Г. Борисовой (Борисова 1995а).

В настоящее время термин «коллокация» нашел широкое применение в корпусной лингвистике, в рамках которой понятие коллокации переосмысливается или упрощается по сравнению с традиционной лингвистикой. Этот подход смело можно назвать статистическим. Во главу угла ставится частота совместной встречаемости, поэтому

[1] Данная работа выполнена при частичной поддержке гранта РФФИ 10-07-00563-а «Создание интегрированной автоматизированной системы для лингвистических исследований».

коллокации в корпусной лингвистике могут быть определены как с*татистически устойчивые словосочетания*. При этом статистически устойчивое сочетание может быть как фразеологизированным, так и свободным. 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.

И именно появление больших репрезентативных корпусов текстов позволяет получить достоверные данные о частоте того или другого сочетания в языке в целом. Высокая величина частоты совместной встречаемости, казалось бы, говорит об устойчивости данного сочетания. Однако этой характеристики недостаточно, чтобы говорить о предпочтительной сочетаемости тех или других слов. Поэтому был выработан целый ряд статистических мер (они получили название «меры ассоциации», или меры ассоциативной связанности, англ. association measures), вычисляющих силу связи между элементами в составе коллокации. В общем случае, эти меры учитывают как частоту совместной встречаемости, так и другие параметры, прежде всего частоту в данном корпусе каждого отдельного элемента.

Тем не менее, одних статистических данных недостаточно. Необходимо ответить на вопрос, каким еще требованиям должны соответствовать такие статистически устойчивые словосочетания.

2. Коллокации под углом зрения статистики

Практически большинство корпусных менеджеров обладают способностью производить подсчеты частот слов или словоформ и частот совместной встречаемости. Существует большое число мер ассоциации, которые основываются на этих данных. Общее количество этих мер исчисляется многими десятками. Значения мер ассоциации можно считать показателями силы синтагматической связи между элементами словосочетаний. Описание наиболее распространенных мер см. (Evert 2004). Чаще других используются MI, t-score и log-likelihood. Некоторые корпусные менеджеры предоставляют возможность вычисления этих мер.

Мера ***MI (mutual information)***, введенная в работе (Church, Hanks 1990), сравнивает зависимые контекстно-связанные частоты с независимыми, как если бы слова появлялись в тексте совершенно случайно:

$$MI(n,c)=log_2 \frac{f(n,c)\times N}{f(n)\times f(c)}, \text{ где}$$

n — ключевое слово (node); c — коллокат (collocate); f(n,c) — частота встречаемости ключевого слова n в паре с коллокатом c; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и слова c в корпусе (тексте); N — общее число словоупотреблений в корпусе (тексте).

Если значение MI (n,c) больше определенного значения (для русского языка часто называется значение 3 и больше), тогда данное сочетание слов можно считать статистически значимым. Для MI (n,c) меньше нуля говорится, что *n* и *c* находятся в отношении дополнительной дистрибуции.

Есть различные модификации этой формулы, различными способами повышающие значение f(n,c) (Oakes 1998; Kilgarriff 2001).

Мера ***t-score*** также учитывает частоту совместной встречаемости ключевого слова и его коллоката, отвечая на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами:

$$t-score=\frac{f(n,c)-\frac{f(n)\times f(c)}{N}}{\sqrt{f(n,c)}}$$

Также достаточно часто применяется мера, известная под названием ***log-likelihood***, или *логарифмическая функция правдоподобия* (Dunning 1993).

$$log\text{-}likelihood = 2\sum_{ij} Oij \times \log\frac{Oij}{Eij}, \text{ где}$$

Oij, Eij — наблюдаемая и ожидаемая частоты (подробнее см. (Evert 2004: 83)).

3. Цель и методы

Цель работы — сравнительный анализ различных мер ассоциации на основе корпусов русского языка. Кроме того, исследуется зависимость результатов (списка коллокаций, полученного на основе одной и той же меры) от текстового материала (тип текста). Работа выполнялась, в основном, на материале корпусов университета г. Лидс, составленных С. А. Шаровым на базе разных подмножеств Национального корпуса русского языка (НКРЯ), корпуса Интернет-текстов и др.

Сервис на сайте университета г. Лидс[2]позволяет выбрать одну из трех мер ассоциации (MI, t-score, log-likelihood) (варианты формул для вычисления выбраны С. А. Шаровым) или их совокупность, указать часть речи коллоката и расстояние между словами. Также имеется возможность проводить поиск коллокатов по лемме или словоформе. Не-

[2] http://corpus.leeds.ac.uk/ruscorpora.html

обходимо отметить, что каждый элемент в корпусе, включая знаки препинания, считается словом. Как следствие, среди результатов оказываются бессмысленные по сути комбинации, например, глаголов со знаками препинания.

Основным материалом исследования послужили 10 частотных глаголов русского языка: *быть*, *сказать*, *мочь*, *говорить*, *знать*, *стать*, *есть*, *хотеть*, *видеть*, *идти*. Эксперимент заключался в нахождении биграмм, одним из компонентов которых является глагол из приведенного списка.

4. Результаты

Результаты поиска коллокаций по корпусу НКРЯ-[3]были сведены в таблицы, представляющие собой объединение коллокаций, полученных на основе трех вышеуказанных мер (см. Табл. 1). Далее были удалены бессмысленные коллокации, т. е. комбинации глаголов со служебными словами и знаками пунктуации. Каждой коллокации был приписан свой ранг.

[3] Подмножество в 50 млн. словоупотреблений

Табл. 1. Часть таблицы результатов для глагола «***говорить***» (левый контекст) (модель Adv+V), отсортированных по мере MI

	Collocation	Joint	Freq1	Rank MI	MI score (7,08–2,14)	Rank LL	LL score (1064,06–2,96)	Rank T-score	T- score (22,79–1,96)
33.	честно говорить	527	2339	1.	7,08	1.	1064,06	101.	1,96
34.	постоянно говорить	62	4158	2.	7,04	14.	40,59	85.	2,26
35.	условно говорить	90	585	3.	6,53	8.	162,73	91.	2,11
36.	обиженно говорить	5	208	4.	6,46	77.	4,37	16.	6,52
37.	грубо говорить	130	988	5.	6,30	6.	224,23	93.	2,10
38.	умело говорить	23	2034	6.	6,20	33.	12,26	70.	2,40
39.	откровенно говорить	139	1203	7.	6,12	5.	230,24	94.	2,09
40.	собственно говорить	333	3114	8.	6,00	2.	538,32	99.	1,97
41.	жалобно говорить	6	481	9.	5,91	94.	3,45	25.	4,96
42.	……								

В столбцах таблицы, кроме самой коллокации, указываются следующие характеристики: *Joint — частота совместной встречаемости, Freq1 — частота коллоката (левый контекст), Rank MI — ранг по мере MI, MI score — значение меры MI, Rank LL — ранг по LL (log-likelihood), LL score — значение LL, Rank T-score — ранг по t-score, T-score — значение t-score.*

Анализ данных Табл. 1 (всего 101 коллокация) показывает, что ранги коллокаций, полученных на основе разных мер, не совпадают. Наиболее отличается мера t-score; меры MI и LL, наоборот, часто демонстрируют близкие ранги для найденных коллокаций (см. Табл. 1, строки 1, 3, 5, 7, 8), что верно и для других глаголов и синтаксических конструкций.

В словарных статьях толковых словарей для глагола «*говорить*» особенно отмечены устойчивые словосочетания, компонентом которых является форма деепричастия «*говоря*». Мы провели поиск биграмм, компонентом которых является именно данная форма слова. Ниже приведена сравнительная таблица для этого случая (см. Табл. 2).

Табл. 2. Частотные данные и меры ассоциации для глагола «***говорить***» (первое значение для леммы /***второе значение курсивом*** для формы деепричастия)

Collocation	Joint	Freq1	MI score	LL score	T score 1,96
искренне говоря	12/***5***	1562	2,94/***4,92***	4,49/***6,11***	2,74/***2,16***
точно говоря	66/***41***	9893	2,64/***5,29***	21,09/***55,31***	2,21/***6,24***
просто говоря	214/***144***	28089	2,19/***5,60***	79,38/***209,98***	2,02/***11,75***
откровенно говоря	139/***104***	1203	6,12/***9,67***	230,24/***299,54***	2,09/***10,19***
честно говоря	527/***429***	2339	7,08/***10,98***	1064,06/***1690,55***	1,96/***22,33***
объективно говоря	7/***6***	502	4,24/***6,82***	4,37/***11,22***	4,16/***2,43***
образный говоря	51/***44***	233	3,00/***10,80***	102,07/***145,01***	2,32/***6,63***
строгий говоря	166/***146***	4239	4,55/***8,34***	184,16/***351,80***	2,08/***12,05***
условно говоря	90/***87***	585	6,53/***10,45***	162,73/***275,55***	2,11/***9,32***
грубо говорить	130/***127***	988	6,30/***10,24***	224,23/***392,55***	2,10/***11,26***
мягко говоря	252/***247***	3916	5,27/***9,22***	341,77/***672,86***	2,01/***15,69***
коротко говоря	267/***265***	6540	4,61/***8,58***	301,73/***662,08***	1,97/***16,24***
собственно говоря	333/***332***	3114	6,00/***9,97***	538,32/***996,77***	1,97/***18,20***
упрощенно говоря	5/***5***	34	3,07/***10,44***	8,92/***15,78***	7,53/***2,23***

Табл. 3. Значения мер ассоциации для коллокаций со словом ***война,*** совпавших с коллокациями словаря Е. Г. Борисовой

Collocation	Joint	Freq1	Freq2	LL score	MI	T-score
вспыхивать война	5	1201		**8,29**	**6,20**	**2,21**
идти война	167	47464		**264,43**	**5,96**	**12,72**
кровопролитный война	6	251		**15,18**	**8,72**	**2,44**
разражаться война	9	881		**18,94**	**7,50**	**2,98**

Табл. 4. Меры ассоциации для коллокаций со словом ***война*** по БАС-17

Collocation	Joint	Freq1	Freq2	LL score	MI	T-score
гражданский война	194	12469		**451,11**	**8,10**	**13,88**
локальный война	8	860		**16,46**	**7,36**	**2,81**
мировой война	154	25171		**285,92**	**6,76**	**12,29**
партизанский война	45	728		**135,77**	**10,09**	**6,70**
холодный война	171	4747		**469,90**	**9,31**	**13,06**

Анализ результатов показывает, что в этом случае коллокации обладают значительно бóльшим числовым значением для всех мер ассоциации. См., например, строки «точно говоря», «откровенно говоря». Из этого можно сделать вывод, что иногда статистические меры для поиска коллокаций следует применять к словоформам, а не к леммам.

Далее мы исследовали зависимость состава и ранжирования списков коллокаций, полученных на основе одной и той же меры MI на разных корпусах текстов русского языка (НКРЯ (117 млн. словоупотреблений) и газетный корпус (70 млн.)), то есть зависимость списка коллокаций от типа текста.

Анализ коллокаций, полученных на этих двух корпусах, показывает, что грубо их можно разбить на две части: присутствующие в обоих корпусах (часто с близкими рангами) и присутствующие только в одном из них. Видимо, это говорит о принадлежности коллокатов, в данном случае, наречий, выданных только по одному из корпусов, к определенному жанру. И действительно, анализ контекстов употребления наречий *пространно, модно, полусерьзно, фигурально* показывает преобладание их в корпусе художественных текстов. Но еще более разительную картину дает сравнений коллокаций, полученных на основе НКРЯ (117 млн. словоупотреблений) и Интернет-корпус (188 млн.), где из 13 первых коллокаций из НКРЯ, отсортированных по мере MI, в Интернет-корпусе присутствует только одна. Это говорит о том, что для разных жанров, возможно, следует применять разные меры.

Также следует учитывать, что разные меры поразному реагируют на частоту слов, образующих коллокацию, и на частоту совместной встречаемости. Так утверждается, что MI чувствительна к низкочастотным словам, а t-score полезна для нахождения высочастотных коллокаций (Evert 2004; Čermák 2006; Kilgarriff 2006).

Мы провели сравнение коллокаций, полученных автоматически на основе разных мер ассоциации, с данными различных словарей (более подробно см. (Хохлова 2008)). Материалом послужили коллокации 19 существительных, не имеющих омонимов (по Малому академическому словарю (МАС) (Словарь русского языка 1981–1984)) и представленных в словаре коллокаций русского языка Е. Г. Борисовой (Борисова 1995b). Исследование проводилось также на базе корпусов русских текстов, созданных в университете г. Лидс. Были проанализированы коллокации на базе газетного корпуса (78 млн. слов).

Результаты запроса для каждого существительного (выявленные коллокации) сравнивались со словарными статьями, приведенными для этих существительных в Словаре коллокаций (Борисова 1995b), в толковых словарях русского языка: БАС-17 (Словарь современного русского языка 1948–1965) и МАС (Словарь русского языка 1981–1984) — и в Словаре синонимов и сходных по смыслу выражений (Абрамов 2006). Приведем некоторые результаты для слова *война* (по первым 100 биграммам для левого контекста) (Таблица 3).

Будем называть коллокации, приведенные в словаре Е. Г. Борисовой и входящие в таблицы, «правильными».

В таблице 4 приведены словосочетания, найденные на слово *война* в БАС-17.

На рис. 1 приводится график для БАС-17 (значения меры MI по оси ординат и ранги биграмм по оси абсцисс). Темным цветом обозначены «правильные» коллокации из словаря Борисовой (ранги 12, 18, 38, 41) и дополнительные коллокации, найденные в БАС-17 (ранги 2, 8, 13, 34, 44).

Мы видим, что «правильные» коллокации Борисовой и устойчивые словосочетания БАС распределены в левой половине шкалы.

На таком же графике для БАС-17 по значениям меры log-likelihood «правильные» коллокации из словаря Борисовой (ранги 5, 35, 43 и 60) и дополнительные коллокации, найденные в БАС-17 (ранги

1, 2, 3, 5, 8), также находятся в левой части графика. Для словосочетаний, найденных по мере t-score, мы также наблюдаем «скученность» «словарных» коллокаций в левой половине графика.

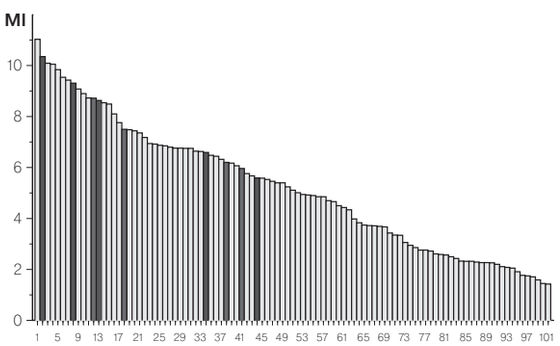

Рис. 1. Значения меры MI для коллокаций со словом ***война***

Для всех полученных сочетаний наблюдается одинаковая тенденция: чем меньше значение меры, тем больше вероятность, что эти словосочетания не зафиксированы как устойчивые в словарях русского языка. Таким образом, можно сказать, что данные о сочетаемости, приведенные в словарях, совпадают с данными, полученными на основе мер ассоциации.

Важным представляется и тот факт, что в результате эксперимента были выделены сочетания, не зафиксированные ни в одном из словарей. Анализ подобных сочетаний показал, что биграммы, находящие на самом верху списка (отсортированного по убыванию по одной из мер), с некоторой долей вероятности оказываются устойчивыми и, следовательно, могут быть внесены в словарь.

Как уже было сказано, поверх статистических критериев должны работать и другие методы, основывающиеся на собственно лингвистических моделях. Данная идея заложена и реализована в известной системе Sketch Engine (Kilgarriff et al. 2004). Она выдает для заданного ключевого слова типичные словосочетания, обусловленные, с одной стороны, синтаксисом, накладывающим ограничение на сочетаемость слов в заданном языке, а с другой стороны, вероятностными закономерностями, связанными с семантикой и языковым узусом. Результат работы программы представлен наиболее устойчивыми словосочетаниями с учетом грамматических (структурных) формул. Однако и с помощью обычных корпусных менеджеров также можно получить похожие результаты (см. Табл. 5, поиск на сайте университета г. Лидс).

Результаты поиска и выдачи коллокаций в таком виде удобны для лексикографов, которые могут выбрать для словарей примеры разного типа, и для лингвистов, изучающих лексику и синтаксис в определенном аспекте.

Табл. 5. Результаты поиска коллокаций с глаголом «***сказать***» (отсортированные по мере LL)

V + Adv (Pred)		**V + N**	
Collocation	**LL**	**Collocation**	**LL**
сказать трудно	56,37	сказать правда	31,68
сказать нельзя	20,33	сказать слово	14,13
сказать точно	18,34	сказать гадость	6,44
сказать вслух	17,65	сказать комплимент	4,73
сказать особо	16,90	сказать неправда	4,60
сказать честно	12,90	сказать тост	3,92
можно сказать	3274,07	хотеть сказать	2550,77
надо сказать	1768,42	хотеться сказать	247,53
нельзя сказать	524,47	успевать сказать	81,72
трудно сказать	518,11	следовать сказать	70,24
точно сказать	183,63	забывать сказать	68,04
тужно сказать	145,33	смочь сказать	47,21

5. Выводы

При сравнении сочетаний, полученных с помощью статистических методов, со словарями наблюдается одинаковая тенденция: чем меньше значение меры, тем больше вероятность, что эти словосочетания не зафиксированы как устойчивые в словарях русского языка, и наоборот. Большинство коллокаций, зафиксированных в словарях, оказывается в верхней части списка, составленного на основе одной из мер ассоциации. Таким образом, можно сказать, что данные об устойчивой сочетаемости, приведенные в словарях, совпадают с данными, полученными на основе мер ассоциации, или, по-другому, что статистические меры ассоциации достаточно хорошо выявляют реально существующие семантико-синтагматические связи.

Сравнительный анализ различных мер ассоциации, проведенный по совокупности всех данных, полученных нами для разных частей речи, показывает следующее.

Мера MI, возможно, дает наилучшие усредненные результаты. Она позволяет выделить устойчивые фразеологизированные словосочетания, а также сочетаниями, где в качестве коллокатов выступают имена собственные, а также низкочастотные специальные термины. К недостаткам использования меры t-score можно отнести то, что она, в первую очередь, выделяет коллокации с очень частотными словами-коллокатами, в частности, со служебными словами. Поэтому для t-score необходимо задавать список стоп-слов, чтобы «отбросить» самые частотные слова, сочетания с которыми неизменно оказываются вверху таблицы: предлоги, местоимения или союзы. Впрочем, это, видимо, справедливо и для других мер. Это следует из наших экспериментов, это же подтверждается и в других публикациях (см., напр., Baroni 2008; Evert 2004; Čermák 2006;

Khokhlova 2009b; Kilgarriff 2006; Křen 2006; Пивоварова 2010; Хохлова 2008).

Возможно, что стоит изучить возможности объединения разных мер, например, ввести величину, равную сумме их рангов (Cvrček 2006).

Остается открытым вопрос, стоит ли учитывать в статистических мерах при поиске коллокаций леммы или словоформы (см. Табл. 2).

Следует также принимать во внимание структурные синтаксические формулы и семантические ограничения, которые лежат в основе коллокаций. Их комбинация со статистическими подходами, по нашему мнению, может дать неплохие результаты (Khokhlova 2009a).

В качестве общего вывода из проведенного исследования мы хотим отметить, что существующий программный инструментарий автоматического выявления коллокаций на основе статистических методов весьма неудовлетворителен — как в части лингвистического обеспечения, так и с точки зрения выходных интерфейсов, и его следует развивать. В первую очередь, важно уметь находить разрывные коллокации со свободным порядком, искать коллокаты не только по леммам, но и по словоформам, искать коллокаты для гнезда опорных однокоренных слов, уметь варьировать размер окна, в котором ищутся коллокаты. Нередко реальные коллокации представляют собой n-граммы, где n больше двух, тогда встает вопрос выбора формул для мер ассоциации для таких словосочетаний. При выборке коллокаций из текста должна производиться обработка знаков препинания и служебных слов, имен собственных и т. п.

Особое значение имеет выдача коллигаций — коллокаций, построенных по определенной синтаксической модели, учет отношения зависимости между элементами коллокаций.

Литература

1. *Абрамов Н. М.* Словарь русских синонимов и сходных по смыслу выражений. М.: 2006.
2. *Ахманова О. С.* Словарь лингвистических терминов, М.: 1966.
3. *Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю.* Словарь глагольной сочетаемости непредметных имен русского языка. М., 2002. (См. // http://dict.ruslang.ru/)
4. *Борисова Е. Г.* Коллокации. Что это такое и как их изучать. М.: 1995. (Борисова 1995a).
5. *Борисова Е. Г.* Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарем ключевых слов. М.: 1995. (Борисова 1995b).
6. *Денисов П. Н., Морковкин В. В.* Словарь сочетаемости слов русского языка. М., 2002.
7. *Иорданская Л. Н., Мельчук И. А.* Смысл и сочетаемость в словаре. М.: 2007.
8. *Кустова Г. И.* Словарь русской идиоматики. Сочетания слов со значением высокой степени. // http://dict.ruslang.ru/
9. *Пивоварова Л. М.* Подводные камни статистических мер (в печати). 2010.
10. *Словарь* русского языка: В 4 т.: 1981–1984, Т. 1–4, Москва. (МАС)
11. *Словарь* современного русского литературного языка: В 17 т.: 1948–1965, Т. 1–17, Москва, Ленинград. (БАС-17).
12. *Хохлова М. В.* Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. Под ред. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасовой. Хельсинки: 2008. С. 343–357.
13. *Atkins S. and Rundell M.* The Oxford Guide to Practical Lexicography. Oxford University Press, 2008.
14. *Baroni M., Evert S.* Statistical methods for corpus exploitation. // In A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook, article 36. Mouton de Gruyter, Berlin: 2008.
15. *Benson M.* (Ed.) The BBI combinatory dictionary of English. Amsterdam: John Benjamins Publishing Co, 1986.
16. *Čermák F.* Metoda zjistovani kolokacni platnosti frekventovanych bigramu pomoci ranku. // Čermák F. (ed.) Kolokace. Ústav Českého národního korpusu, Praha: 2006. P. 94–105.
17. *Church K., Hanks P.* Word association norms, mutual information, and lexicography // Computational Linguistics, 1996, № 16(1), P. 22–29.
18. *Crowther J., Dignen S. & Lea D.* (Eds.). Oxford Collocations Dictionary for Students of English. Oxford: Oxford University Press, 2002.
19. *Cvrček V.* Metoda zjišťování kolokační platnosti frekventovaných bigramů pomocí ranku. // Čermák F. (ed.) Kolokace. Ústav Českého národního korpusu, Praha: 2006. P. 36–55.
20. *Dunning T.* Accurate Methods for the Statistics of Surprise and Coincidence. // Computational Linguistics. 1993. Volume 19, №1, P. 61–74.
21. *Evert S.* The Statistics of Word Cooccurences Word Pairs and Collocations. PhD thesis. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart: 2004.
22. *Khokhlova M.* Applying Word Sketches to Russian. // In Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, P. 91–99. (Khokhlova 2009a)
23. *Khokhlova M., Zakharov V.* Statistical collocability of Russian verbs // After Half a Century of Slavonic Natural Language Processing. Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.). Brno, 2009. P. 105–112. (Khokhlova 2009b)
24. *Kilgarriff A.* Collocationality (and how to measure it) // Proceedings of the Euralex International Congress. Torino, 2006.
25. *Kilgarriff A.* Web as Corpus. // Proceedings of Corpus Linguistics. Lancaster, UK. 2001.
26. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* The Sketch Engine // Proceedings of the Eleventh EURALEX International Congress. Lorient, 2004. P. 105–116.
27. *Kjellmer G.* A dictionary of English collocations: based on the Brown corpus : in three volumes. Oxford; New York: Clarendon Press: Oxford University Press, 1994.
28. *Křen M.* Kolokacni miry a cestina: srovnani na datech ČNK. // Čermák F. (ed.) Kolokace. Ústav Českého národního korpusu, Praha: 2006. P. 223–248.
29. *Krishnamurthy R. & Keith B.* 2006. Collocations Encyclopedia of Language & Linguistics. Oxford: Elsevier. P. 596–600.
30. *Oakes M.* Statistics for Corpus Linguistics. Edingurgh: 1998.
31. *Sinclair J.* Collins COBUILD English collocations on CD-ROM: Harper Collins, 1995.
32. *Sinclair J.* Corpus, concordance, collocation. Oxford: Oxford University Press, 1991.