Пивоварова Л.М., Ягунова Е.В. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов (предварительные наблюдения) // Материалы Симпозиума "Терминология и знание" (Москва, 21-22 мая 2010 г.) М., 2010

Пивоварова Л.М., Ягунова Е.В. (Санкт-Петербург, Россия)

ИЗВЛЕЧЕНИЕ И КЛАССИФИКАЦИЯ ТЕРМИНОЛОГИЧЕСКИХ КОЛЛОКАЦИЙ НА МАТЕРИАЛЕ ЛИНГВИСТИЧЕСКИХ НАУЧНЫХ ТЕКСТОВ (ПРЕДВАРИТЕЛЬНЫЕ НАБЛЮДЕНИЯ)

We present first step of research aimed to find formal approach for an identification of collection's terminology: subject domain and key words. We use collocation extraction techniques for this purpose. We demonstrate that two statistical measures MI and t-score are useful for collection description, though MI is better for subject domain detection. The topic homogeneity of collection corresponds with homogeneity of the extracted collocations set. The collocations that exist in all the collection (on in the majority of texts) may be detected by high value of t-score measure. We believe, that set of extracted collocations may be considered a «current vocabulary».

1. Введение. Постановка задачи

В настоящее время проводится исследование, целью которого является изучение возможности выделения формальных признаков, необходимых для определения предметной области коллекций текстов и ключевых слов, описывающих рассматриваемые коллекции. Анализируются коллекции (или корпуса) научных и новостных стилей (ср. работы [Ягунова, Пивоварова 2010; Ягунова 2010]). Научные тексты мы рассматриваем в пределах общей предметной области «Теоретическая и прикладная лингвистика». В данном докладе мы хотели бы сосредоточиться на идеях автоматического извлечение информации об основных темах лингвистических научных текстов; реализация этих идей неотделима от задач выделения терминов, наиболее важных для сопоставляемых коллекций.

В научных текстах доминирует информативная функция, для этих текстов значимы (и определимы) тематические сферы употребления. Для текстов научного стиля в целом характерна более жесткая смысловая и коммуникативная структурированность текста (композиция, структура фрейма). Большинство интересующих нас терминов этой предметной области оказываются неоднословными. При выделении коллокаций в научном тексте с помощью статистических мер, мы можем ограничить задачу исследования коллокаций вообще и сосредоточиться на исследовании терминологических коллокаций, прежде всего, биграмм. Коллокации понимаются нами как неслучайное сочетание двух и более лексических единиц, характерное как для языка в целом (текстов любого типа), так и оп-

ределенного типа текстов (или даже (под)выборки текстов). Для выделения использование различных статистических мер, позволяющих автоматически выделить из текстов коллокации и ранжировать их по степени устойчивости в соответствии со значениями выбираемых мер (подробнее об использованных нами мерах см. п.2).

Последнее время все чаще появляются работы, в которых рассматриваются пути решения задач автоматического выделения терминологических коллокаций (неоднословных терминов), чаще всего для индексирования документов в задачах информационного поиска или пополнения словарей интеллектуальных систем (см., например, в работах [Добров и др. 2003; Браславский, Соколов 2006]).

Чрезвычайно актуальным статистический метод может быть тогда, когда изучаются процессы становления новой предметной области, изменения терминологии (особенно при сосуществования разных научных парадигм, каждая из которых может использовать свой терминологический аппарат). Множество терминологических коллокаций, выделяемое на заданной коллекции научных текстов, характеризует узкую предметную область (темы и подтемы) этой коллекции[1].

В литературе обсуждается вопрос об (оперативных) единицах анализа текста и о единицах словаря: «текущего» словаря, который учитывает подстройку адресата – или системы автоматического анализа – под особенности конкретных текстов. Подстройка может включать в себя анализ особенностей текстов и – обязательно – собственно подстройку, т.е. формирование «текущего словаря», позволяющего оптимальным образом учитывать особенности единиц анализа конкретных текстов (подробнее см. в работе [Ягунова 2008]). В зависимости от этих параметров выбираются единицы анализа, разные как по объему, так и по уровням анализа:

- лексема и/или словоформа, биграмма или n-грамма (единицами которых могут быть как лексемы, так и словоформы);
- единица, функционирующая как слово (состоящее из одного или более орфографических слов) или единица, соответствующая устойчивым конструкциям (в том числе и предикативным) вплоть до высказывания.

Вероятно, процесс выделения наиболее значимых коллокаций непосредственно соотносится с задачей классификации предметных областей (что особенно актуально для

[1] Дополнительная подзадача может включать в себя выделение коллокаций, характеризующих стиль и тип текста (например, *зависит от*, *отличие от*, *дает возможность*, (*в*) *первую очередь*, (*в*) *свою очередь* и т.д.). Она не является предметом анализа в рамках данного доклада, в целом же в рамках нашего исследования коллокаций, характеризующих научный текст, эта задача дополняет задачу выделения терминологических коллокаций, их совместное решение позволяет охарактеризовать рассматриваемые коллекции научных текстов с точки зрения и содержательной информации, и стилистических особенностей.

междисциплинарных областей научного знания) и формированием стратифицированных «текущих словарей». Этот процесс может быть в свою очередь соотнесен со следующими друг за другом этапами подстройки под все более сужающуюся предметную область, например:

- научные тексты,
  - лингвистические научные тексты,
    - научные тексты предметной области «Теоретическая и прикладная лингвистика»,
      - научные тексты предметной области «Корпусная лингвистика».

В основу данной работы положены следующие **гипотезы:**

- используемые в работе статистические меры (MI и t-score, подробнее см. ниже, п. 2) позволяют охарактеризовать предметную область рассматриваемых коллекций;
- предметную область текстов коллекции лучше характеризуют коллокации, выделяемые с помощью меры MI;
- степень тематической однородности коллекции научных текстов соотносится с однородностью множества выделяемых коллокаций;
- коллокации, общие для **всех** (или **подавляющего большинства)** текстов коллекции, характеризуются высокими значениями меры t-score.

На данном – предварительном – этапе исследования мы ограничились рассмотрением лишь двухсловных коллокаций (биграмм).

2. Материал и методика

В качестве основного материала использовались две коллекция научных текстов:

1. Монотематическая коллекция материалов конференции «Корпусная лингвистика» 2004-2008 года[2]. Объем коллекции составляет около 220000 «токенов» - словоупотреблений и знаков препинания.

2. Коллекция материалов международной конференции «Диалог» «Компьютерная лингвистика и интеллектуальные технологии» за 2003-2009 годы. Объем коллекции составляет около 2,5 миллионов «токенов» – словоупотреблений и знаков препинания[3].

[2] Пользуясь случаем, хотим поблагодарить кафедру «Математической лингвистики» филологического факультета СПбГУ и лично О.А. Митрофанову за любезно предоставленную нам для работы коллекцию текстов.

Выбор материала определялся желанием взять две коллекции, сопоставимые по тематике при том, что одна включает в себя тексты гораздо более широкой тематики, чем другая. Таким образом, коллекция материалов конференции «Корпусная лингвистика» рассматривается как монотематическая, (точнее – гораздо более узкотематическая, чем коллекция конференции «Диалог»).

Морфологическая разметка коллекций осуществлялась В.В. Бочаровым[4] при помощи свободно распространяемого программного обеспечения АОТ (www.aot.ru). Для разметки использовался, в первую очередь, модуль морфологической анализа; модуль синтаксического анализа использовался для частичного снятия морфологической омонимии. В тех случаях, когда полностью снять омонимию не удавалось (по приблизительным оценкам — около 6% случаев), для анализа использовалась первая из предложенных анализатором лемм, т.е. неоднозначность разбора просто игнорировалась. Такое решение было принято в связи с тем, что на сегодняшний день остается не вполне ясным, как учитывать неоднозначность разбора в используемых нами статистических мерах MI и t-score. При выделении коллокаций учитывалась пунктуация: рассматривались любые последовательности слов в тексте, не разделенных знаками препинания.

На данном этапе нами использовались две меры *MI* и t-score (см. об этих мерах подробнее в обзорах [Stubbs 1995; Хохлова 2008]).

MI (mutual information, коэффициент взаимной информации) сравнивает зависимые контекстно-связанные частоты с независимыми, как если бы слова появлялись в тексте совершенно случайно:

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

где

MI – объем информации; n – n-е ключевое слово; c – коллокат;

f(n,c) – абсолютная частота встречаемости ключевого слова *n* в паре с коллокатом *c*;

f(n), f(c) – абсолютные частоты ключевого слова *n* и слова *c* в корпусе;

N – общее число словоформ в корпусе.

С точки зрения теории вероятности, мера MI является способом проверить независимость появления двух слов в тексте — если слова полностью независимы, то вероятность их совместного появления равна произведению вероятностей появления каждого из них, т.е. произведению частот (использование абсолютных частот вместо относительных

---

[3] Предварительная подготовка текстов коллекции была проведена Д.Грановским. Пользуясь случаем, хотим поблагодарить Д.Грановского и выразить надежду на продолжение совместной работы.

[4] Пользуясь случаем, выражаем благодарность В.В. Бочарову и надеемся на дельнейшее плодотворное сотрудничество.

увеличивает значение MI для всех коллокаций в корпусе на константу, однако не меняет ее вероятностного смысла). Из определения видно, что мера MI зависит от размера корпуса — чем больше исследуемый корпус, тем выше в среднем получаемые по нему значения MI. Это свойство, видимо, должно отражать большую степень доверия к данным полученным на материале большего корпуса. Однако в настоящем исследовании мера MI используется как средство ранжировать коллокации внутри одного корпуса по степени их связности – сравнение между коллекциями осуществляется лишь по рангу, но не по значению выделенных биграмм.

Другим недостатком меры MI, который отмечают многие исследователи (в том числе в работах [Stubbs, 2008; Manning, Schutze 2002]), является ее свойство завышать значимость редких словосочетаний. Чем более редки слова, образующие коллокацию, тем выше будет для них значение MI, что делает данную меру совершенно «беззащитной» перед опечатками, иностранными словами и другим информационным шумом, который неизбежен в большой коллекции. Поэтому для данной меры используется порог отсечения по частоте – в данной работе мы рассматривали только те биграммы, которые встретились в коллекции не менее сорока раз для коллекции «Диалог» и не менее 16-ти раз для корпусной коллекции (данные значения подбирались интуитивно, однако высокие пороги отсечения определялись самой постановкой задачи: нашей целью является выделить наиболее значимые, характерные для данной коллекции словосочетания, т.е. акцент делается на точности, а не на полноте).

Необходимо отметить, что как правило при подсчете меры MI порядок слов внутри коллокации не учитывается — данная мера отражает взаимозависимость двух лексем, но не значимость конкретной коллокации. В данной работе, однако, ***учитывался порядок*** коллокатов: мера MI подсчитывалась в отдельности для каждой конкретной пары лексем.

Другой мерой, которая использовалась в данном исследовании, стала мера *t-score,* которая учитывает частоту совместной встречаемости целевого слова и его коллоката, отвечая на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами. Мера t-score, расчитывается по формуле (условные обозначения здесь приняты те же, что и выше для MI):

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

Данная мера используется гораздо реже, чем мера MI, поскольку она является лишь несколько модифицированным ранжированием коллокаций по частоте. Очевидно, что значение данной меры тем выше, чем выше частота коллокации в коллекции. Хотя данная

мера содержит коррекционный компонент — вычитание деленного на размер коллекции произведения частот коллокатов, однако эта поправка отражается лишь на самых частотных словах. Это свойство часто делает данную меру малопригодной для поиска терминологических словосочетаний и для этой цели она, как правило, не используется. Мера t-score, в отличие от MI, не преувеличивает значимость редких коллокаций, поэтому ее часто используют без порогов отсечения. Тем не менее, мы использовали для t-score те же пороги отсечения, что и для MI, чтобы в обоих случаях работать с одним и тем же множеством коллокаций, и также учитывали порядок коллокатов внутри биграммы.

3. Результаты. Обсуждение результатов

3.1. Биграммы, выделяемые с помощью меры MI

Для каждой из двух коллекций было получено по два списка биграмм: список лексемных биграмм и список словоформных биграмм. Биграммы в списках упорядочены по значению меры. Излагаемые в докладе результаты относятся к верхушке этих списков (90 биграмм с наиболее высокими значениями меры).

Первоначальные списки (см. на примере табл. 1 и 2) включали в себя как терминологические биграммы, так и биграммы, характеризующие стиль текстов рассматриваемых коллекций. Мы полагаем весьма показательным сопоставление биграмм, выявленных для лексем и/или для словоформ.

На материале новостных текстов был проведен предварительный сопоставительный анализ (1) списка биграмм, выделяемых для лексем (но не словоформ), (2) списка биграмм, выделяемых для словоформ (но не лексем) и (3) списка биграмм, выделяемых и для лексем, и для словоформ[5] (подробнее см. статью [Ягунова, Пивоварова 2010]).

В список 1 попадают составные номинации, характеризуемые максимальной свободой (максимальным разнообразием, минимальной ограниченностью) набора выполняемых ими в предложении семантико-синтаксических ролей. Примеры биграмм первого списка (число обозначает п.п., каждая единица сочетания приведена в нормализованном виде (словарной форме)): 9 *винительный падеж*, 17 *именительный падеж*, 24 *актуальный членение*, 29 *инструментальный среда*.

Биграммы второго списка, как правило, относятся к номинации в определенной синтаксической позиции. Примеры биграмм этого списка (число обозначает п.п.): 10 *речевой акт*, 50 *речевых актов*, 19 *именная группа*, 65 *именных групп*, 27 *коммуникативного акта*, 62 *коммуникативных актов*, 77 *просодических характеристик*, 78 *прошедшего вре-*

[5] Во всех случаях речь идет о сравнении первых ста биграмм из списка

мени, 74 *речевого сигнала*. Кроме того, биграммы этого подкласса могут относиться к части целостной номинации, например., сочетание *речевых актов* часто является частью триграммы «теории речевых актов»[6]. У биграмм третьего списка (т.е. пересечения списков для биграмм и словоформ) обычно наиболее простая структура, как правило, в этих структурах нет ни закрепленности, ни противоречий между смысловыми, лексическими и синтаксическими связями. Биграммы этого класса занимают в текущем словарном составе некое **промежуточное место** между биграммами класса «1» и биграммами класса «2».

На наш взгляд, этот третий класс – биграммы, выделяющиеся и для лексем, и для словоформ – наиболее информативен для решения задач, поставленных в данном исследовании. Таким образом, мы выделяем наиболее информационно-нагруженные и точные сочетания, характеризующие данную коллекцию (см. табл. 1 (1а) и 2 (2а)). Для простоты восприятия в таблицах биграммы списка представлены в виде сочетаний словоформ (соответствующей словоформной биграмме).

Таблица 1. Биграммы (MI-score), выделяющиеся
и для лексем, и для словоформ.
Материал конференции «Диалог»

п.п.	Биграммы	
1	ударном	Слоге
2	концептуальных	Графов
4	внешним	посессором
5	оперативной	Памяти
8	вокального	жеста
14	*крайней*	*мере*
16	XIX	Века
17	лингвистического	процессора
21	положение	Дел
22	*первую*	*очередь*
25	картине	Дира
26	множественного	числа
28	интеллектуальные	технологии
30	корпусная	лингвистика
33	отглагольных	существительных
37	знаки	препинания
38	педагогической	коммуникации
42	основного	Тона
46	машинного	перевода
61	устойчивых	словосочетаний
63	точки	Зрения
70	*меньшей*	*Мере*
72	*вряд*	*Ли*

[6] Такие более длинные коллокации, как правило, оказываются в верхней части списка, полученного с использованием обобщенных мер MI и t-score для более длинных, чем биграммы, последовательностей слов; сравнению n-грамм различной длины будут посвящены дальнейшие части нашего исследования.

73	предметной	области
85	*вплоть*	*До*

Курсивом в таблице выше выделены сочетания, которые были удалены на этапе выделения терминологических коллокаций с использованием морфологического фильтра. Подчеркиванием выделены те сочетания, которые на основании формальных критериев должны были быть ошибочно отнесены к терминологическим.

Таблица 1а. Терминологические биграммы (MI-score),
выделяющиеся и для лексем, и для словоформ.
Материал конференции «Диалог»

п.п.	Биграммы	
1	Ударном	Слоге
2	Концептуальных	Графов
4	Внешним	Посессором
5	Оперативной	Памяти
8	Вокального	Жеста
16	XIX	Века
17	Лингвистического	Процессора
25	Картине	Мира
26	Множественного	Числа
28	Интеллектуальные	Технологии
30	Корпусная	Лингвистика
33	Отглагольных	Существительных
37	Знаки	Препинания
38	Педагогической	Коммуникации
42	Основного	Тона
46	Машинного	Перевода
61	Устойчивых	Словосочетаний
73	Предметной	Области

Таблица 2. Биграммы (MI-score),
выделяющиеся и для лексем, и для словоформ.
Материал конференции «Корпусная лингвистика»

п.п.	Биграммы	
2	*Наш*	*взгляд*
3	*(по) крайней*	*мере*
4	Речевой	деятельности
5	Художественной	литературы
7	*Первую*	*очередь*
9	Общим	объемом
11	Корпусная	лингвистика
13	Имена	собственные
15	Математической	лингвистики

16	Словарной	статьи
17	*Свою*	*очередь*
18	Предметной	области
19	Машинного	перевода
20	Точки	зрения
22	*За*	*счет*
24	*Речь*	*идет*
25	*Прежде*	*всего*
26	Большое	количество
28	Настоящее	время
31	*Представляет*	*собой*
32	*Млн*	*словоупотреблений*
34	*Другой*	*стороны*
35	Семантических	состояний
36	*Одной*	*стороны*
37	*Таким*	*образом*
40	Разрешения	неоднозначности
41	Английский	язык
43	*Кроме*	*того*
47	Национальный	корпус
48	Грамматических	категорий
52	Устная	речь
54	База	данных
58	*Во*	*многих*
61	Лексических	единиц
62	*Дает*	*возможность*
63	*Зависит*	*От*
64	*Отличие*	*От*
65	Русский	язык
67	Корпусные	данные
68	*Отличается*	*От*
71	*Зависимости*	*От*
72	*Работы*	*Над*
79	Частей	речи
80	*Во*	*всех*
84	*При*	*помощи*
86	Морфологической	разметки
87	*Говорить*	*О*

Таблица 2а. Терминологические биграммы (MI-score),
выделяющиеся и для лексем, и для словоформ.
Материал конференции «Корпусная лингвистика»

п.п	Биграммы	
4	Речевой	деятельности
5	Художественной	литературы
9	Общим	объемом
11	Корпусная	лингвистика
13	Имена	собственные
15	Математической	лингвистики
16	Словарной	статьи
18	Предметной	области

19	Машинного	перевода
26	Большое	количество
35	Семантических	состояний
40	Разрешения	неоднозначности
41	Английский	язык
47	Национальный	корпус
48	Грамматических	категорий
52	Устная	речь
54	База	данных
61	Лексических	единиц
65	Русский	язык
67	Корпусные	данные
79	Частей	речи
86	Морфологической	разметки

На наш взгляд из сопоставления двух списков – двух множеств ведущих терминологических биграмм, полученных с помощью меры MI (табл. 1а и табл. 2а) – можно сделать, вывод о высокой общности тематики, с одной стороны, и, с другой стороны, о том, что степень тематической однородности коллекции научных текстов соотносится с однородностью множества выделяемых коллокаций. Список наиболее устойчивых терминологических сочетаний, полученный для текстов конференции «Корпусная лингвистика», дает представлении об обобщенном содержании этой коллекции. Этого списка достаточно, чтобы получить предварительную информацию о наиболее важных объектах исследования, материале, методах, результатах (ср. с исследованием по ключевым словам для текстов этой коллекции [Ягунова 2010]). Вернее было бы сказать, что представление об обобщенном содержании коллекции «Корпусной лингвистике» будет более полным, чем о содержании коллекции конференции «Диалог».

3.2. Биграммы, выделяемые с помощью меры t-score

Как уже было сказано, данная мера используется гораздо реже, чем мера MI, поскольку она является лишь несколько модифицированным ранжированием коллокаций по частоте. Обычно она считается малопригодной для поиска терминологических словосочетаний, не используясь для этой цели.

Однако в процессе работы над новостными коллекциями мы обнаружили, что эта мера оказывается полезна при решении задачи о выделении тех единиц, которые характеризуют **все** (или **подавляющее большинство**) текстов коллекции. Основная масса таких сочетаний характеризует скорее особенности стиля текстов коллекции, впрочем, используя минимальный морфологический фильтр из списков t-score-коллокаций, мы могли выделить те сочетания, которые могут рассматриваться как терминологические. Таким обра-

зом был получен список терминологических биграмм, общих для **всех** (или **подавляющего большинства**) текстов рассматриваемых коллекций (см. табл. 3 и 4).

Таблица 3. Терминологические биграммы (t-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Диалог»

п.п.	лексемные биграммы		П.п.	словоформные биграммы	
2	Русский	Язык	3	русского	языка
			12	русском	языке
63	Предметный	область	49	предметной	области

Таблица 4. Терминологические биграммы (t-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Корпусная лингвистика»

п.п.	лексемные биграммы		п.п.	словоформные биграммы	
2	Русский	Язык	4	русского	Языка
4	Корпус	Текст	66	корпус	Текстов
			22	корпуса	Текстов
21	Часть	Речь	54	части	Речи
29	Английский	Язык	42	английского	Языка
55	Машинный	Перевод	46	машинного	Перевода
63	Предметный	Область	74	предметной	Области
79	Лексический	Единица	77	лексических	Единиц

Сопоставление полученных списков приводит нас к следующим выводам:

1. Тематика конференции Диалог настолько широка, что на основании общих терминологических сочетаний мы могли бы сделать вывод лишь о том, что, как правило, в качестве основного материала исследований выступает русский язык, а также, что предметной области часто уделяется внимание.

2. Представляемые на «Корпусной конференции» исследования чаще всего ориентированы на русский или английский язык. В качестве материала (и/или объекта исследования) выступают корпуса, что лексическим единицам (частям речи) уделяется особое внимание. Что многие исследования ориентированы на решение вопросов машинного перевода. Таким образом, наши выводы согласуются с традиционной тематикой корпусных исследований, что отражено в наборе «общих» терминологических сочетаний.

4. Заключение

Несмотря на то, что данное исследование можно считать сугубо предварительным, основные гипотезы на рассматриваемом материале подтвердились:

- используемые в работе статистические меры (MI и t-score) позволяют охарактеризовать предметную область рассматриваемых коллекций;
- степень тематической однородности коллекции научных текстов соотносится с однородностью множества выделяемых коллокаций;
- коллокации, **общие** для **всех** (или **подавляющего большинства**) текстов коллекции, характеризуются высокими значениями меры t-score.

Мы не ставили перед собой задачу практически востребованного метода извлечения ***всех*** терминов или тестирования разных методик (см., например, работу [Браславский, Соколов 2006]). Тем более не предполагалось использование констрастивного «общеязыкового» корпуса или общенаучного корпуса. Нашей задачей было изучение возможности выделения формальных признаков, необходимых для определения предметной области коллекций текстов и ключевых слов, описывающих рассматриваемые коллекции; формирование наборов значимых для коллекции терминологических коллокаций и выделение общих для текстов коллекции терминологических коллокаций. Полагаем, что основные пути решения этой задачи были намечены. Конечно, на настоящий момент получены ответы далеко не на все вопросы. На наш взгляд одним из следующих будет вопрос о том, будет ли полученный список (или вернее, списки) упоминаемыми во введении «текущими» словарями, полезными и необходимыми для анализа конкретных коллекций? Мы полагаем, что это весьма плодотворная гипотеза, которая ждет своей проверки и последующих уточнений.

Литература

1. Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста //Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) /Под ред. Н.И. Лауфер, А. С. Нариньяни, В. П. Селегея. – М., 2006. – С. 88 – 94
2. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - RCDL2003. – Санкт-Петербург, 2003. – С. 201 – 210
3. 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.

4. Ягунова Е.В. Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). – Пермь: Издательство Пермского государственного университета, 2008. – 395 с.

5. Ягунова Е.В. Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов //Материалы IV Международного конгресса исследователей русского языка «Русский язык: исторические судьбы и современность». – М ., 2010. – С. 533-534

6. Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов //Научно-техническая информация . Сер.2. Информационные процессы и системы.– № 6 – 2010. – С. 30 – 40.

7. Stubbs M. Collocations and semantic profiles: on the case of the trouble with quantitative studies // Functions of language 2(1), 23-55, Benjamins, 1995.

8. Manning C., Schutze H. Collocations //Manning C., Schutze H. Foundations of Statictical Natural Language Processing,. – MIT Press. Cambridge, 1999. – P. 151 – 189.