Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Сб. НТИ, Сер.2, №6. М., 2010.

Гипертекстовый вариант статьи размещён здесь:
http://www.webground.su/services.php?param=priroda_collac&part=priroda_collac.htm

Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов

Е.В. Ягунова, Л.М. Пивоварова

1. Введение. Цели исследования. Постановка задачи

Коллокации понимаются нами как неслучайное сочетание двух и более лексических единиц, характерное как для языка в целом (текстов любого типа), так и определенного типа текстов (или даже (под)выборки текстов). В литературе достаточно часто встречается понимание лингвистами коллокаций как несвободных сочетаний, не относящихся к идиомам, когда, с одной стороны, ключевое слово этих сочетаний может появляться в контексте разных языковых единиц, с другой стороны эти единицы (т.е. контекст ключевого слова) можно перечислить в виде закрытого списка (ср., напр., работы Л. Н. Иорданской, И. А. Мельчука и их последователей по исследованию лексических функций и моделей управления[1]). Однако этот список отражает, главным образом, интуицию исследователя и лишь в некоторой степени может быть соотносим с исследованием тех особенностей, которые не просто заложены в языке (всех текстах на этом языке), но в существенной степени зависят от типа рассматриваемых текстов. Альтернативой интуитивному методу можно считать 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). Для нас этот последний метод является единственно приемлемым, т.к. в нашем исследовании рассматриваются большие массивы текстов разных функциональных стилей и предметных областей, а список потенциальных коллокаций для них принципиально не задан, поскольку этот список является отражением тех языковых и экстралингвистических характеристик, которые заложены в анализируемых текстах., и выявление которых является конечной целью данного исследования

Поскольку текст есть структурированная последовательность единиц разных уровней, сложносоставные подструктуры текста (коллокации) выступают важным объектом при исследовании процедур анализа (и синтеза) текста. Таким образом, создается возможность исследования структурных единиц текста разных языковых – и текстовых – уровней и их роль в процедурах анализа и синтеза речи (текстов). Результаты такого исследования могут быть востребованы специалистами в самых разных областях: теоретическая и прикладная лингвистика, автоматический анализ текста. Более того, формализация методов анализа и синтеза речи, отражающая реальные стратегии восприятия и порождения речи человеком, является одной из наиболее фундаментальных проблем междисциплинарной области исследования восприятия речи человеком (т.е. задачей для лингвистов, физиологов, психологов, работающих в этом направлении).

В литературе обсуждается вопрос об (оперативных) единицах анализа текста и о единицах словаря. Вернее будет говорить не о словаре вообще, а о «текущем» словаре, который учитывает подстройку адресата под особенности конкретных текстов. Термин «текущий словарь» применительно к проблемам восприятия текста человеком был введен в (Венцов, Касевич 1994). В случае изучения восприятия речи человеком адресат очевиден. В случае с автоматическим анализом текстов, по-видимому, в качестве адресата следует иметь в виду как систему анализа («посредник» адресата), так собственно адресата (пользователя этой системы). Несмотря на возможные существенные различия работы человека и систем автоматического анализа, необходимость формирования и подстройки текущего словаря является их общей особенностью - как известно, большинство эффективных систем анализа и синтеза текстов работает на ограниченной предметной области и стиле текстов (ср., напр., процедуры подключения разнообразных предметных словарей при машинном переводе). 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. Вариативность выбора этих единиц напрямую соотносится с неединственностью стратегий анализа речи, задаваемых, в частности задачей и типом текстов (подробнее см. 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.

В результате развития корпусной лингвистики в последнее время стали появляться частотные словари, учитывающие различие таких единиц как лексема и/или словоформа в зависимости от функционального стиля (или типа) текстов, напр., «Статистический словарь русской газеты (1990 гг.)» (Шайкевич и др. 1998) и «Новый частотный словарь русской лексики»[2], созданный на базе НКРЯ (Ляшевская, Шаров 2008). Из известных нам лингвистических словарей лишь в «Статистическом словаре русской газеты» представлены все биграммы «с частотой 3 и более (включая самые тривиальные грамматические сочетания)» (Шайкевич и др. 1998), выделяемые на основании абсолютной частоты встречаемости[3]. Сложность задачи выделения «значимой» лексики для определенного типа текстов (напр., «значимой газетно-новостной лексики») наглядно иллюстрирует «Новый частотный словарь русской лексики»[4].

Однако, при всей важности и нужности названных словарей их безусловно недостаточно для исследования потенциальных «текущих словарей», отражающих закономерности разных типов текстов. Полагаем, что анализ коллокаций, которые входят в состав научных и новостных текстов, в наибольшей степени отражает специфику появления «особых» единиц, характеризующих предметную область и тип текста, что, вероятно, непосредственно соотносится с формированием «текущих словарей». Этот процесс может быть в свою очередь соотнесен со следующими друг за другом этапами подстройки:

- тексты определенного функционального стиля,
  - тексты определенного источника,
    - тексты определенной предметной области,

- однородная выборка текстов определенных источников и предметной области,

  - и т.д.

Какова же возможная природа коллокаций? Чаще всего коллокации рассматривают тогда, когда пытаются образом выделить фразеологизмы (единицы с элементами фразеологизации) или наиболее жесткие модели управления, а также разнообразные варианты текстовой реализации лексических функций (при использовании статистических методов для выделения коллокаций можно говорить о совпадении двух пониманий термина «коллокация»). К коллокациям могут обращаться тогда, когда исследуют дискурсивные слова, составные служебные слова, сложные номинации и т.д.[5] (напр., Кобрицов и др.,2005; Degand & Bestgen, 2003). Такого рода работы проводятся обычно на материале художественных и/или новостных текстов (последние выступают источником во многом благодаря легкому доступу к такого рода текстам). Но на наш взгляд большинство такого рода работ не нацелены на изучение специфики именно художественных и/или новостных текстов, и тех ограничений, которые накладывает тип текста на функционирование языковых единиц.

Наряду с этим все чаще появляются работы, в которых рассматриваются пути решения задач выделения терминологических коллокаций (неоднословных терминов), чаще всего для индексирования документов в задачах информационного поиска или пополнения словарей интеллектуальных систем (см., напр., напр. Добров и др. 2003; Браславский, Соколов 2006).

Новостные тексты представляют собой один из наиболее проблематичных типов текстов. Актуальные для СМИ темы возникают и исчезают – таким образом, основная предметная область таких текстов, с одной стороны, очевидно выделяется (это ведущие темы), но, с другой стороны, она нестабильна – в отличие от предметной области научных текстов.

Другое отличие новостных текстов состоит в степени выраженности ведущей функции языка и степени структурированности текста. 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). В новостных текстах – как и в научных – доминирует информативная функция языка, однако может реализовываться и функция воздействия на адресата. Таким образом, они могут быть неоднородны в отношении предметной области, функции и структуры.

Какие номинации можно ожидать при анализе новостных текстов? Полагаем, что для новостных текстов будет характерно выделение тех сложных номинаций, которые характеризуют конкретную выборку (или, напр., подборку новостей). Таким образом, вероятно, лидирующими будут те сложные номинации, которые информационно значимы либо для новостных текстов в целом, либо для данной конкретной выборки. При этом они могут не иметь общеязыковой значимости. Например, такие традиционно рассматриваемые сложные номинации как *железная дорога*, *белый медведь*, *холодная война* могут не попадать в список сложных номинаций для новостных текстов (т.е. они, вероятно, окажутся в списке лишь в том случае, если будет высокой их информационная значимость для конкретной выборки текстов)[6]. Сходным образом, по-видимому, будет обстоять дело и для неоднословных терминов (в отличие от употребления терминов в текстах научного функционального стиля). Термины в новостных текстах несут

собственно терминологический характер лишь в том случае, если предметная область новостного текста совпадает с предметной областью научных (или деловых) текстов. Сравнительно часто в новостных текстах присутствуют номинации, исходно носящие терминологический характер, но давно и прочно вошедшие в языковую практику (напр., *стихийное бедствие*).

Полагаем, что наряду с коллокациями, характеризующими предметную область рассматриваемых текстов, новостные тексты должны иметь и общие родовые черты, которые выражаются в использовании служебных слов и устойчивых оборотов, не связанных с тематикой конкретных наборов текстов. Одна из целей данного исследования – изучить возможности разделения «тематических» и «стилистических» коллокаций при помощи различных статистических мер.

Таким образом, в данной работе верификации подлежали следующие ***гипотезы***:

1. используемые в работе статистические меры (MI и t-score, подробнее см. п.2) позволяют охарактеризовать предметную область и стилистические особенности новостных текстов;

2. списки коллокаций, полученных с помощью MI и t-score, принципиально различны:

    a. коллокации, выделяемые с помощью MI, позволяет определять наименования объектов, термины, сложные номинации, отражающие предметную область,

    b. критерий t-score направлен на выделение «общеязыковых устойчивых сочетаний» (производных служебных слов, дискурсивных слов) и «устойчивых конструкций», где и те, и другие характеризуют стилистические особенности новостных текстов;

3. коллокации, выделяемые для монотематической коллекции научных текстов (взятая для сопоставления с новостными) характеризуются большей однородностью:

    a. коллокации, выделяемые с помощью MI, точно определяют предметную область,

    b. коллокации, выделяемые с помощью t-score дают представление о наборе «общеязыковых» устойчивых сочетаний (или, скорее «общенаучных» или «общих для рассматриваемой коллекции»).

2. Материал и методика

В качестве основного материала использовалась коллекция текстов портала www.lenta.ru с апреля по декабрь 2009. Общий объем проанализированных текстов: более 66000000 «токенов» (словоупотреблением и знаков препинания). Морфологическая разметка коллекции осуществлялась В.В. Бочаровым[7] 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 учитывались знаки препинания: рассматривались любые последовательности слов в тексте, не разделенных знаками препинания.

Как уже было сказано, на данном этапе нами использовались две меры *MI* и t-score (см. об этих мерах подробнее в обзорах Khokhlova 2008 и Хохлова 2008)[8].

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

$$MI=log_2 \qquad ,$$

где

MI – объем информации;

n – 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 увеличивает значение MI для всех коллокаций в корпусе на константу, однако не меняет ее вероятностного смысла). Также 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. В то же время, такая «дискриминация» по размеру корпуса затрудняет сравнение значений мер, полученных на разных корпусах. Этот вопрос еще будет подробно рассматриваться в дальнейшей работе; в настоящем исследовании мера MI используется как средство ранжировать коллокации внутри одного корпуса по степени их связности.

Другим недостатком меры MI, который отмечают многие исследователи (в том числе Stubbs, 2008; Manning, Schutze 2002 и др.), 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 (данное значение подбиралось интуитивно и 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 ключевого 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):

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. Stubbs (2005) показывает (на примере английского языка), что значение меры t-score для знаменательных слов примерно равно $\sqrt{f(n, c)}$ и лишь для служебных заметно меньше этого значения. Это свойство делает данную меру малопригодной для поиска терминологических словосочетаний и для этой цели она, как правило, не используется. Также из вышесказанного следует, что мера t-score, в отличие от MI, не преувеличивает значимость редких коллокаций и не нуждается в использовании 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.

Одной из основных гипотез, лежащей в основе данной работы, является принципиальное различие двух рассматриваемых статистических мер:

1. MI наилучшим образом позволяет автоматически выделять неоднословные целостности, соответствующие сложным номинациям;

2. t-score, напротив, лучше работает при выделении стилистических особенностей, производных служебных слов и «устойчивых конструкций» (на данном материале устойчивыми конструкциями являются, прежде всего, «газетные штампы»).

В формате данной статьи мы ограничились биграммами[9] и сосредоточились на коллокациях, выделяемых с помощью меры MI (t-score используется только для сопоставления).

Традиционно считается, что с помощью меры MI хорошо выделяются имена собственные и низкочастотные специальные термины, что можно использовать при решении задач информационного поиска. Слова, у которых MI-score принимает наибольшую величину, обладают ограниченной сочетаемостью. В качестве материала используется однородный массив текстов представительного объема: рассматривается лишь один новостной ресурс и лишь один год. Можно ли считать, что выделяемые коллокации задают предметную область? Именно это является **основной гипотезой,** верифицируемой на данном материале.

В данной работе ставятся также следующие задачи:

- какова природа коллокаций, автоматически выделяемых с использованием меры MI (в т.ч. в сравнении с теми, которые были выделены с помощью меры t-score);
- какие типы коллокаций выделяются наилучшим vs. наихудшим образом;
- что представляет типичная структура подобной коллокации

Как было сказано выше, в исследовании рассматривались только биграммы, которые встретились в коллекции более 40 раз. В текстах портала www.lenta.ru за 2009 нашлось 11141 таких биграмм – сочетаний контактных лексем – и 8490 биграмм – сочетаний контактных словоформ. Из списка были удалены биграммы, включающие слово(-а), написанные латиницей. Затем биграммы упорядочивались по убыванию значения меры MI или t-score. Первичному содержательному анализу подлежали первые 100 биграмм из получившихся списков, которые получили в работе содержательную классификации и интерпретацию.

Вопрос о выборе первичной лексической единицы анализа – лексемы и/или словоформы – для русского языка (как языка с развитой морфологией) всегда решается неоднозначно. Он зависит от целей исследования коллокаций, от типа текстов и от многих дополнительных факторов. Мы в своей работе анализировали обе эти единицы как отражающие разные аспекты и уровни лексико-грамматической информации об исследуемых единицах. Для словоформ использовались те же формулы и пороги отсечения, что и для лексем.

3. Результаты

3.1. Биграммы, выделяемые с помощью меры MI

Среди первых 100 биграмм лексемных биграмм, выделяемые с помощью меры MI, большинство составляли имена собственные: 43 наименования лица, 17 наименований объектов (главным образом, организаций), 10 географических наименований. Среди этих биграмм были выделено 25 устойчивых сочетаний, условно разделенных на сочетания терминологического и общеязыкового характера[10] (приблизительно поровну: 13 и 12 соответственно). Для тех биграмм, в которых могут быть установлены синтаксические отношения между коллокатами, ведущее место занимают биграммы с атрибутивной связью (31 биграмма) и лишь 6 биграмм имеют генетивную связь (как дополнительный способ выражения атрибутивного значения).

Среди первых 100 биграмм из словоформ встретились повторения лишь двух номинаций: (Саудовская Аравия (1пп.) и Саудовской Аравии (66 пп.)), Барак Абама (Бараком Обамой (26 пп.) Бараку Обаме (38 пп.) и печально известного названия ночного

клуба (Хромой лошади (23 пп.) и Хромая лошадь (59 пп.)). Большая часть и этих биграмм представляли собой имена собственные, однако их доля существенно ниже, чем в случае лексемных биграмм. Лишь 20 из этих биграмм – это наименования лица, 23 – наименования объекта (или часть этого наименования, напр., *Женской теннисной* из *Женской теннисной ассоциации*), 16 – географических наименований (или их части). Среди биграмм из словоформ доля сочетаний, претендующих на устойчивость, выше, чем для лексемных биграмм: 36 сочетания, 25 из них нами условно признаны терминологическими[11], а 11 – имеют скорее общеязыковой характер. Словоформы как единицы биграмм демонстрируют морфологически оформленные синтаксические отношения. В анализируемой части этих биграмм 56 связано атрибутивной связью и лишь 2 биграммы имеют генетивную связь (как дополнительный способ выражения атрибутивного значения); кроме того, 6 биграмм содержат два прилагательных (являются компонентом атрибутивного комплекса).

Для примера в таблице 1 приведена верхняя часть списка рассматриваемых биграмм.

Таблица 1. Биграммы с наиболее высокими значениями меры MI[12]

лексемы		пп.	словоформы		пп.
Бритни	**Спирс**	1	*Саудовская*	*Аравия*	1
Эльвира	**Набиуллина**	2	**Эльвира**	**Набиуллина**	2
Ле	**Бурже**	3	парниковых	газов	3
Пан	ги	4	**мысе**	**Канаверал**	4
Курманбек	Бакиев	5	Соединенных	Штатов	5
Алишер	Усманов	6	Женской	теннисной	6
Бенедикт	XVI	7	**дельты**	**Нигера**	7
Усейн	Болт	8	Ред	Уингс	8
Лионель	**Месси**	9	**Норильского**	**никеля**	9
мыс	**Канаверал**	10	кредитном	портфеле	10
бин	**Ладен**	11	Палестинской	автономии	11
сердечный	приступ	12	**Бритни**	**Спирс**	12
Осама	бин	13	встречную	полосу	13
Норильский	**Никель**	14	Нижнем	Новгороде	14
дельта	**Нигер**	15	ценным	бумагам	15
стихийное	бедствие	16	беспилотных	летательных	16
АК	*Барс*	17	Пан	Ги	17
тротиловый	**эквивалент**	18	федеральную	трассу	18
тройская	**унция**	19	Млечного	Пути	19
Ролан	Гаррос	20	**тройскую**	**унцию**	20
лампа	накаливание	21	непосредственной	близости	21
Радован	Караджич	22	желтую	карточку	22
полезный	ископаемый	23	**Ле**	**Бурже**	23
Джонни	Депп	24	однополых	браков	24
Фидель	Кастро	25	*Хромой*	*лошади*	25
Дель	*Потро*	26	Бараком	Обамой	26
Дель	*Торо*	27	Континентальной	хоккейной	27
долина	Сват	28	**тротиловом**	**эквиваленте**	28
Арбат	**Престиж**	29	**адронного**	**коллайдера**	29
Саддам	Хусейн	30	годовом	исчислении	30
РАО	*ЕЭС*	31	одиночном	разряде	31

Салават	**Юлаев**	32	Федеральному	собранию	32
симфонический	оркестр	33	**Арбат**	**Престиж**	33
Арсений	*Яценюк*	34	подводных	лодок	34
кровный	месть	35	**Салават**	**Юлаев**	35
голубой	**фишка**	36	**Лионель**	**Месси**	36
Рафаэль	Надаль	37	правоохранительным	органам	37
Римма	Салонен	38	Бараку	Обаме	38
адронный	**коллайдер**	39	мужском	одиночном	39
круглый	стол	40	собственному	желанию	40
Гарри	Поттер	41	Питтсбург	Пингвинс	41
Роберто	Мичелетти	42	**голубых**	**фишек**	42
заработный	плата	43	**бин**	**Ладена**	43
боснийский	серб	44	спиртных	напитков	44
Чен	Ир	45	*Невский*	*экспресс*	45

Можно ли считать, что рассматриваемые биграммы задают предметную область? Интуитивно кажется, что ответ на вопрос будет положительным. Но есть ли особые маркеры, выделяющие события именно этого года? Какова полнота и точность описания главных тем? Это отдельные вопросы, требующие отдельного исследования, включающего эксперименты по отдельным выборкам (в частности выборок по разным периодам (и источникам), а также кластеров, выделяемых на основании разных критериев[13]). Пока что нет возможности ни подтвердить, ни опровергнуть это предположение.

Однако, по-видимому, в целом ряде случаев применение критерия MI, выделяющего сочетания по принципу ограниченной сочетаемости, выдает список именно тех устойчивых сочетаний, о которых обычно принято говорить. Приведем примеры такого рода устойчивых сочетаний (полужирный шрифт выделяет общие биграммы для двух списков):

- **для лексем**: *СЕРДЕЧНЫЙ ПРИСТУП, СТИХИЙНЫЙ БЕДСТВИЕ, ЛАМПА НАКАЛИВАНИЕ, ПОЛЕЗНЫЙ ИСКОПАЕМОЕ, СИМФОНИЧЕСКИЙ ОРКЕСТР, КРОВНЫЙ МЕСТЬ, КРУГЛЫЙ СТОЛ, ЗАРАБОТНЫЙ ПЛАТА,* ***НЕПОСРЕДСТВЕННЫЙ БЛИЗОСТЬ****,* ***СЕЛЬСКИЙ ХОЗЯЙСТВО****, КОРОТКИЙ ЗАМЫКАНИЕ*;

- **для словоформ**: *встречную полосу,* ***непосредственной близости****, однополых браков, спиртных напитков, лесных пожаров, сенсорным экраном, политическое убежище,* ***сельского хозяйства****, профессиональном ринге, высоком разрешении, сексуальных меньшинств*.

Среди словоформных биграмм гораздо больше видна терминологическая направленность, в то же время они настолько актуальны и общеизвестны, что допустимо отнесение подобных биграмм и к общеязыковым.

Таким образом, рассматриваемые устойчивые сочетания характеризуются устойчивостью именно в общеязыковом масштабе, некоторые из них ближе к терминологическим сочетаниям или канцеляризмам, ну а в целом их набор характеризует преимущественно новостные тексты.

Таблица 2. Пересечение между биграммами для лексем и для словоформ для анализируемой первой сотни[14]

пп.	биграммы для лексем		пп.	биграммы для словоформ	
1	БРИТНИ	СПИРС	1	Британи	Спирс
2	ЭЛЬВИРА	НАБИУЛЛИН	2	Эльвира	Набиуллина
3	ЛЕ	БУРЖЕ	23	Ле	Бурже
9	ЛИОНЕЛЬ	МЕССИ	36	Лионель	Месси
10	МЫС	КАНАВЕРАЛ	4	мысе	Канаверал
11	БИН	ЛАДЕН	43	бин	Ладена
14	НОРИЛЬСКИЙ	НИКЕЛЬ	9	Норильского	никеля
15	ДЕЛЬТА	НИГЕР	7	дельты	Нигера
17	АК	БАРС	50	Ак	Барс
18	ТРОТИЛОВЫЙ	ЭКВИВАЛЕНТ	28	тротиловом	эквиваленте
19	ТРОЙСКИЙ	УНЦИЯ	20	тройскую	унцию
20	РОЛАН	ГАРРОС	70	Ролан	Гаррос
26	ДЕЛЬ	ПОТРО	49	дель	Торо
27	ДЕЛЬ	ТОРО	87	дель	Потро
29	АРБАТ	ПРЕСТИЖ	33	Арбат	Престиж
31	РАО	ЕЭС	96	РАО	ЕЭС
32	САЛАВАТ	ЮЛАЕВ	35	Салават	Юлаев
34	АРСЕНИЙ	ЯЦЕНЮК	51	Арсений	Яценюк
36	ГОЛУБОЙ	ФИШКА	42	голубых	фишек
39	АДРОННЫЙ	КОЛЛАЙДЕР	29	адронного	коллайдера
52	РЕН	ТВ	94	РЕН	ТВ
53	ТУРНИРНЫЙ	ТАБЛИЦА	69	турнирной	таблице
54	НЕПОСРЕДСТВЕННЫЙ	БЛИЗОСТЬ	21	непосредственной	близости
57	АРКАДИЙ	ДВОРКОВИЧ	92	Аркадий	Дворкович
59	ГЕРМАН	ГРЕФ	83	Герман	Греф
60	ДА	ВИНЧИ	63	да	Винчи
61	АДЕНСКИЙ	ЗАЛИВ	53	Аденском	заливе
62	КОНТИНЕНТАЛЬНЫЙ	ХОККЕЙНЫЙ	27	Континентальной	хоккейной
63	САУДОВСКИЙ	АРАВИЯ	1	Саудовская	Аравия
			66	Саудовской	Аравии
65	БЕСПИЛОТНЫЙ	ЛЕТАТЕЛЬНЫЙ	16	беспилотных	летательных
67	ХРОМОЙ	ЛОШАДЬ	25	Хромой	лошади
			59	Хромая	лошадь
75	ВЕРА	ЗВОНАРЕВ	88	Вера	Звонарева
88	НЕВСКИЙ	ЭКСПРЕСС	45	Невский	экспресс
			95	Невского	экспресса
92	ОБОГАЩЕНИЕ	УРАН	85	обогащению	урана
94	СЕЛЬСКИЙ	ХОЗЯЙСТВО	73	сельского	хозяйства
98	ПАЛЕСТИНСКИЙ	АВТОНОМИЯ	11	Палестинской	автономии

Таблица 3. Примеры биграммы для лексем, не нашедшие соответствия для биграмм из словоформ (для анализируемой первой сотни):

п.п.	биграммы для лексем	
5	КУРМАНБЕК	БАКИЕВ
6	АЛИШЕР	УСМАНОВ
7	БЕНЕДИКТ	XVI
8	УСЕЙН	БОЛТ
12	СЕРДЕЧНЫЙ	ПРИСТУП
13	ОСАМА	БИН
16	СТИХИЙНЫЙ	БЕДСТВИЕ
21	ЛАМПА	НАКАЛИВАНИЕ
22	РАДОВАН	КАРАДЖИЧ
23	ПОЛЕЗНЫЙ	ИСКОПАЕМОЕ
24	ДЖОННИ	ДЕПП
25	ФИДЕЛЬ	КАСТРО
28	ДОЛИНА	СВАТ
30	САДДАМ	ХУСЕЙН
33	СИМФОНИЧЕСКИЙ	ОРКЕСТР
35	КРОВНЫЙ	МЕСТЬ
37	РАФАЭЛЬ	НАДАЛЬ
38	РИММА	САЛОНЕН
40	КРУГЛЫЙ	СТОЛ
41	ГАРРИ	ПОТТЕР
42	РОБЕРТО	МИЧЕЛЕТТИ
43	ЗАРАБОТНЫЙ	ПЛАТА
44	БОСНИЙСКИЙ	СЕРБ
45	ЧЕН	ИР

Мы полагаем, что сопоставление биграмм, выявленных для лексем и/или для словоформ, весьма показательно.

Прежде всего, следует обратить внимание на то, что во всех трёх случаях выделяются составные номинации, являющиеся лексически устойчивыми (иными словами, состоящие из двух слов, отношения между которыми характеризуются лексической устойчивостью) [15].

1. Биграммы, выделяемые для лексем (но не словоформ), по-видимому, имеют более «объемную» структуру, они чаще всего соответствуют номинациям, имеющим в текстах разные синтаксические и смысловые связи.

В класс 1 попадают составные номинации, характеризуемые максимальной свободой (максимальным разнообразием, минимальной ограниченностью) набора выполняемых ими в предложении семантико-синтаксических ролей. Показательна высокая процентная доля, которую имеют в этом классе наименования лиц (по-видимому, для многих наименований лиц особо характерна высокая степень разнообразия набора семантико-синтаксических ролей, в которых они выступают). Для сочетаний, входящих в этот класс, попытка ранжировать семантико-синтаксические роли по степени употребительности, разумеется, приведёт к тому, что среди них выделятся более употребительные и менее

употребительные, но максимально характерная для такого сочетания роль будет для него лишь **слегка** более употребительной, чем остальные возможные для него роли.

2. Биграммы, выделяемые для словоформ (но не лексем). Среди них можно выделить два подкласса – «2а» и «2б».

2а. Биграммы этого подкласса, как правило, относятся к номинации в определенной синтаксической позиции (напр., *встречную полосу, Нижнем Новгороде*);

2б. Биграммы этого подкласса могут относиться к части целостной номинации, тогда на уровне уже более целостной номинации и происходит пересечение n-грамм для словоформ и для лексем: напр., *Женской теннисной*: *Женской теннисной ассоциации* (пп.6), *ЖЕНСКИЙ ТЕННИСНЫЙ АССОЦИАЦИЯ* (пп.39), *рейтинге Женской теннисной ассоциации* (пп.4)[16];

Сходство между случаями «2а» и «2б» состоит в том, что в обоих случаях некоторая составная номинация резко тяготеет к выполнению некоторой типичной (излюбленной) для неё семантико-синтаксической роли (то есть «излюбленная» роль для этой номинации оказывается **гораздо** употребительнее остальных возможных для неё ролей). Такое тяготение является частным проявлением более общего закона тяготения номинативных единиц некоторого грамматико-семантического разряда к выполнению некоторой семантико-синтаксической функции, типичной для единиц этого разряда (в том числе и для простых, т.е. однословных, номинаций); о действии такого закона (применительно к простым номинативным единицам, т.е. словам) писали, в частности, Н.Д.Арутюнова, Г.А.Золотова, В.Г.Гак, Ю.Д.Апресян.

Однако, с другой стороны, между случаями «2а» и «2б» налицо существенное различие. Оно состоит в том, что в случае «2б» данная составная (двухсловная) номинация входит в состав некоторого более крупного (трёхсловного или даже более протяжённого) лексически устойчивого сочетания номинаций, тогда как в случае «2а» сочетание данной лексической номинации с её внешним контекстом (точнее, с её непосредственным соседом на синтагматической оси) является не устойчивым, а свободным.

3. Наиболее простая структура у биграмм, выделяющихся и для словоформ, и для лексем, в них, как правило, т.к. в этих структурах нет противоречий между смысловыми, лексическими и синтаксическими связями (см. табл. 1 и 2).

Биграммы этого класса попадают в него по двум разным причинам. Соответственно, внутри него можно выделить два подкласса – «3а» и «3б».

(3а) Сочетания, у которых тоже статистически вырисовывается «излюбленная» синтаксическая роль, однако она противопоставлена остальным возможным для этого сочетания ролям не столь резко, как это было в типе «2», но и не слегка (как это было в классе «1»), а лишь **умеренно**. Иначе говоря, сочетания этого класса занимают в текущем словарном составе некое **промежуточное место** между сочетаниями класса «1» и сочетаниями класса «2».

(3б) Причина попадания в класс «3» может быть и в отсутствии формальной морфологической оформленности. В класс «3б» могут попадать сочетания, состоящие из двух неизменяемых слов (напр., *РАО ЕЭС*, *Бритни Спирс*, *Ле Бурже*). В таких сочетаниях

ни один из членов не содержит в себе морфологического показателя выполняемой им синтаксической роли.

3.2. Биграммы, выделяемые с помощью меры t-score

Биграммы, выделяемые с помощью меры t-score, с позиций этой работы наиболее легко интерпретируемы. В 80% случаев (анализируя первую сотню) мы наблюдаем пересечение списка словоформных и лексемных биграмм (ср. табл. 4, в ней «ГОД В» является единственным примером отсутствия названного пересечения).

Данная мера позволяет выделять высокочастотные коллокации (в частности, коллокации с высокочастотыми компонентами – прежде всего, предлогами). Она эффективна при поиске «общеязыковых устойчивых сочетаний», вернее, при поиске того, что может рассматриваться как устойчивое сочетание для данной коллекции. В случае с однородной новостной коллекцией, эта мера описывает стилистических особенностей данной коллекции, независимо от конкретной тематики сообщений. Выделяемые биграммы относятся к указанию источников информации (напр., *по словам, со ссылкой, РИА Новости)*, место и время (*в течение, во время, в России*).

Сравнительно многие из рассматриваемых биграмм принято рассматривать как единое слово (напр., составные служебные и дискурсивные слова *в течение, в качестве, может быть*[17]) Интересно, однако, что наряду с ожидаемыми общеязыковыми устойчивыми сочетаниями в списках присутствуют те единицы, которые можно назвать «собственно общеновостными устойчивыми сочетаниями»: напр., *РИА Новости*, *миллион долларов*, *миллион рублей, ПО ДАННЫЕ, КАК СООБЩАТЬ, СО ССЫЛКА****[18]*** (ср. табл. 4).

Таблица 4. Биграммы с наиболее высокими значениями меры t-score

ОБ	ЭТО	об	этом
ОДИН	ИЗ	по	словам
ПО	СЛОВО	а	также
А	ТАКЖЕ	со	ссылкой
ПО	ДАННЫЕ	ссылкой	на
ССЫЛКА	НА	по	данным
СО	ССЫЛКА	кроме	того
В	РЕЗУЛЬТАТ	РИА	Новости
КРОМЕ	ТОТ	этом	сообщает
РИА	НОВОСТЬ	при	этом
В	ЧАСТНОСТЬ	в	том
ЭТО	СООБЩАТЬ	в	России
МИЛЛИОН	ДОЛЛАР	во	время
В	РОССИЯ	пока	не
МИЛЛИАРД	ДОЛЛАР	о	том
ВО	ВРЕМЯ	в	результате
ПРИ	ЭТО	настоящее	время
В	КОТОРЫЙ	миллионов	долларов

КАК	СООБЩАТЬ	связи	с
О	ТОМ	сообщает	РИА
В	ХОД	в	результате
В	ТОТ	в	частности
В	СВОЙ	миллиарда	долларов
ПОКА	НЕ	как	сообщает
ГОД	**В**	том	числе
СВЯЗЬ	С	на	сайте
В	МОСКВА	в	ходе
В	КОНЕЦ	стало	известно

4. Заключение

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

Сопоставительный анализ данных подтверждает, что коллокации, выделяемые с помощью MI, отражают предметную область[19]. Информативным при этом оказывается сравнение коллокаций, получаемых для словоформ и для лексем, в результате чего вырисовывается объемная картина, учитывающая смысловые, лексические и синтаксические связи между коллокатами.

Также подтверждена гипотеза о принципиальном различии списков коллокаций, полученных с помощью MI и t-score:

- MI наилучшим образом позволяет выделять наименования объектов, термины сложные номинации;

- t-score, напротив, лучше работает при выделении «общеязыковых устойчивых сочетаний» (производных служебных слов, дискурсивных слов) и «устойчивых конструкций», где и те, и другие характеризуют именно стилистические особенности текстов рассматриваемого типа (в данном случае – новостных текстов).

Для того чтобы определить основные особенности именно новостных текстов, мы, в частности, сравнили результаты с аналогичными данными, полученными на материале монотематической коллекции текстов научного стиля: материалов конференции «Корпусная лингвистика» 2004-2008 года[20] (подробнее см. Ягунова, Пивоварова 2010). Положения, лежащие в основе рассматриваемых задач, подтвердились. В частности:

- коллокации с максимальным значением MI позволяют определить предметную область, что для научных текстов оказывается более простой задачей (многие из них носят терминологический характер)

    - ср. биграммы **для лексем**: *КОРПУСНОЙ ЛИНГВИСТИКА, ПРЕДСТАВЛЯТЬ СЕБЯ, ИМЯ СОБСТВЕННЫЙ, НАСТОЯЩИЙ ВРЕМЯ, ЗАВИСЕТЬ ОТ, ИМЕННОЙ ГРУППА, СЛОВАРНЫЙ СТАТЬЯ, СВОЙ ОЧЕРЕДЬ, СНЯТИЕ НЕОДНОЗНАЧНОСТЬ, ПРИКЛАДНОЙ ЛИНГВИСТИКА*; для **словоформ**: *контекстной предсказуемости, наш взгляд, крайней мере, речевой деятельности, художественной литературы, XX века, первую очередь, что*

касается, общим объемом, имен собственных, корпусная лингвистика, вряд ли, имена собственные, данный момент, математической лингвистики, словарной статьи, свою очередь, предметной области, машинного перевода, точки зрения;

- коллокации с максимальным значением t-score (и для лексем, и для словоформ) дают представление о наборе «общеязыковых устойчивых сочетаний» (или, скорее «общенаучных» или «общелингвистических»): составных слов (напр., *ТАКОЙ ОБРАЗ* или *таким образом*), конструкций, прежде всего, предложно-падежных (напр., *В КОРПУС* или *в корпусе*) и номинаций, общих для рассматриваемой коллекции (напр., *РУССКИЙ ЯЗЫК* или *русский язык*, *КОРПУС ТЕКСТ* или *корпус текстов*), где и те и другие характеризуют особенности текстов именно рассматриваемой коллекции (типа)

    - ср. биграммы **для лексем**: *В КОРПУС, РУССКИЙ ЯЗЫК, И Т* (из *и т.д.*), *КОРПУС ТЕКСТ, А ТАКЖЕ, И В, В ТЕКСТ, ТАКОЙ ОБРАЗ, В ТОТ, МОЧЬ БЫТЬ, ОДИН ИЗ, В ЭТОТ, ТАК И, ПРИ ЭТО, ТОТ ЖЕ, НА ОСНОВА, НЕ ТОЛЬКО, СЛОВО В*; для **словоформ**: *и т* (из *и т.д.*), *может быть, а также, русского языка, в том, в корпусе, так и, не только, таким образом, и др, точки зрения, на основе, но и, могут быть, в тексте, корпуса текстов*.

Таким образом, коллокации, выделяемые с помощью наиболее традиционной меры MI, дают представление о предметной области текстов. Однако, если для научных текстов таким образом сравнительно легко выделяются основные термины, то для новостных коллекций эти коллокации представляют собой крайне неоднородное множество: наименования объектов (лиц, подразделений, географических объектов), термины и общеязыковые устойчивые сочетания.

Мы не ставили перед собой задачу практически востребованного метода извлечения терминов или тестирования разных методик (см., напр., Браславский, Соколов 2006). Тем более не предполагалось использование констрастивного «общеязыкового» корпуса.

На данном этапе нас в большей степени интересовал вопрос о природе коллокаций и тех особенностях текстов, которые они отражают. Ведь автоматически получаемые списки коллокаций и последующий их ручной анализ являются для нас возможностью исследовать те языковые и экстралингвистические характеристики, которые заложены в анализируемых текстах. В частности, будет ли полученный список (или вернее, (под)списки) упоминаемыми во введении «текущими» словарями, полезными и необходимыми для анализа конкретных коллекций (и соответственно подобранных типов текстов)? Каковы особенности поведения биграмм в зависимости от рассматриваемой единицы (лексемы vs. словоформы)? Эти и многие другие вопросы требуют дальнейшего экспериментального изучения и теоретического осмысления. Мы не претендуем на решение вопроса о природе коллокаций, но – на примере биграмм – считаем необходимым поставить этот вопрос, имеющий как теоретическое, так и прикладное значение.

Литература:

Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю. Словарь глагольной сочетаемости непредметных имен русского языка М., 2008 http://dict.ruslang.ru/abstr_noun.php

Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текса // 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.: Изд-во РГГУ, 2006.

Венцов А. В., Касевич В. Б. Проблемы восприятия речи. СПб., 1994.

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

Иорданская Л. Н., Мельчук И. А.. Смысл и сочетаемость в словаре. М.: Языки славянских культур, 2007

Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Поверхностные фильтры для разрешения семантической омонимии в текстовом корпусе // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005" (Звенигород, 1-6 июня, 2005 г.)/ Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. - М.: Наука, 2005.

Кустова Г. И. Словарь русской идиоматики. Сочетания слов со значением высокой степени М., 2008 http://dict.ruslang.ru/magn.php

Ляшевская О. Н., Шаров С. А. Новый частотный словарь русской лексики 2008 http://dict.ruslang.ru/freq.php

Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. Статистический словарь русской газеты (1990 гг.) М., 1998

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

Ягунова Е.В. Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь, 2008.

Ягунова Е.В., Пивоварова Л.М. Извлечение и классификация коллокаций на материале научных текстов. Предварительные наблюдения. СПб., 2010 (в печати)

Degand L., Bestgen Y.Towards automatic retrieval of idioms in French newspaper corpora // Literary and Linguistic Computing, 18, 2003, 249-259

Iordanskaja, L., Paperno, S.: A Russian-English Collocational Dictionary of the Human Body, Columbus/Ohio 1996

Khokhlova M. Extracting Collocations in Russian: Statistics vs. Dictionary // JADT 2008: actes des 9es Journées Internationales d'Analyse Statistique des Données Textuelles, Lyon, 12-14 mars 2008 : Proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12-14, 2008 (editors : Serge Heiden, Bénédicte Pincemin). P. 613–624.

Petrovic S., Snajder J., Basic B.D., Kolar M. Comparison of collocation extraction for document indexing // Journal of Computing and information technology – CIT 14, 2006, 4, 321-327

Stubbs M. Collocations and semantic profiles: om the case of the trouble with quantitative studies.Functions of language 2:11, 23-55, Benjamins, 1995.

Manning C., Schutze H. Collocations // Manning C., Schutze H. Foundations of Statictical Natural Language Processing, 2002, pp.151-189

Rayson, Paul & Roger Garside (2000). Comparing corpora using frequency profiling // Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000. P. 1-6.

---

[1] См. подробнее в (Иорданская, Мельчук 2007; Iordanskaja, Paperno 1996); сейчас такие работы ведутся на основе Национального корпуса русского языка (НКРЯ), в частности, представленные на http://dict.ruslang.ru/ (Кустова 2008; Бирюк и др. 2008).

[2] http://dict.ruslang.ru/freq.php

[3] Во введении к словарю перечисляется несколько десятков единиц, которые войдут в печатный вариант третьего тома словаря. Остается лишь с нетерпеньем ждать выхода этого тома.

[4] Значимая лексика в этом словаре определялась на основе сравнения подкорпуса публицистики и остального корпуса; «в качестве метрики был использован критерий отношения правдоподобия как (Rayson & Garside 2000). Приведем верхнюю часть представленного в словаре списка

(а) упорядоченного по значению логарифма правдоподобия: *страна, театр, наш, новый, советский, компания, который, по, русский, российский, военный, мы, первый, рынок, этот, о, время, работа*;

(б) упорядоченного по частоте в газетном корпусе *по, но, весь, этот, мы, который, свой, о, один, до, другой, время, уже, наш, самый, чтобы, при, очень*. – уже из этих двух списков очевидно различие в результатах, получаемых на основе только частотных характеристик и более сложных статистических мер.

[5] Напр., *золотая молодежь*, *на самом деле*, *в течение, в качестве*, *в связи с* и т.д.

[6] В отличие от текстов литературно-художественного функционального стиля.

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

[8] Пользуясь случаем, выражаем благодарность М.В.Хохловой за консультации.

[9] Хотя обе меры можно обобщить для любого числа коллокатов исследовании (на эту тему см. напр., Petrovic et al 2006; нами проводились исследования коллокаций, включающих от двух до пяти коллокатов.

[10] Деление на сочетания терминологического и общеязыкового характера для новостных текстов довольно условно, т.к. многие номинации, исходно носящие терминологический характер, давно и прочно вошли в общеязыковую практику. Для его анализа желательно формирование тематически однородных выборок текстов, для таких выборок возможно разделить терминологические и нетерминологические варианты использования рассматриваемых сочетаний.

[11] Интересны примеры с биграммами наподобие «темная материя», которые в рассматриваемой новостной коллекции относятся к термину из области физики; в целом, в новостных текстах более частотны термины из области экономики, социологии и криминалистики (соответствующие актуальным темам новостей).

[12] Шрифтовое выделение относится к пересечению лексемных и словоформных биграмм (полужиным – в рамках данной таблицы, курсив – с подчеркиванием в пределах рассматриваемой первой сотни).

[13] В ближайшее время планируется исследование 12 подвыборок рассматриваемой коллекции, соответствующих месяцам 2009 года, что позволит проанализировать динамику тем; а далее – коллекции того же портала за период с 2005-го по 2009 год.

[14] Для удобства рассмотрения лексемы даются прописными, а словоформы строчными буквами.

[15] Пользуясь случаем, выражаем благодарность С.А.Крылову за обсуждение фрагментов статьи на этапе ее написания.

[16] Даже атрибутивное сочетание «федеральную трассу» попадает в эту группу: *ФЕДЕРАЛЬНЫЙ ТРАССА А-114 ВОЛОГДА-НОВЫЙ ЛАДОГА* (пп.1)

[17]Ср. единицы в Корпусном словаре неоднословных лексических единиц (оборотов) на базе НКРЯ http://www.ruscorpora.ru/obgrams.html

[18] Это, очевидно, составные части более длинных выражений «*как сообщает корреспондент*», «*по данным агенства*», «*со ссылкой на*», которые оказываются наверху списка при вычислении t-score для триграм

[19] Отсутствие сравнения полученных данных с генеральным массивом делают невозможным доказательство полноты представленности предметной области, однако это не было задачей данной статьи.

[20] Методика лемматизации, получения биграмм и их анализа полностью аналогична выше описанной. Объем коллекции составляет около 220000 «токенов» - словоупотреблений и знаков препинания.