Е. В. Ягунова, Л. М. Пивоварова ОТ КОЛЛОКАЦИЙ К КОНСТРУКЦИЯМ // Русский язык: конструкционные и лексико-семантические подходы / Отв. ред. С.С.Сай. СПб, 2011 (ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН. Отв. редактор / Н. Н. Казанский)

Е. В. Ягунова, Л. М. Пивоварова,

ОТ КОЛЛОКАЦИЙ К КОНСТРУКЦИЯМ

1. Введение

1.1. Терминология и теоретические предпосылки

Название статьи само по себе содержит некоторую провокацию: названные в нем существительные являются терминами с очень размытым значением. В зависимости от той или иной научной парадигмы изменяется трактовка того, что такое «коллокация» и что такое «конструкция».

При восприятии и порождении (обработке) текста неизбежно используются единицы разной разного масштаба, разной степени связанности и разных уровней иерархии. Эти единицы «задаются» характеристиками языка и контекста, предпочтение тех иных единиц имеет ярко выраженную вероятностную природу. В качестве такого рода оперативных единиц могут выступать как синтаксические, так и лексические единицы (под последними понимаются разнообразные обороты, единицы, эквивалентные слову и т.д. – см., напр., Рогожникова 2005, словарь оборотов www.ruscorpora.ru/obgrams.html).

Однако начнем с попытки разобраться в вопросах терминологии.

В современной лингвистике, ориентированной, с одной стороны, на функциональность и антропоцентричность описания, а с другой стороны – на возможности корпусной лингвистики, уже практически очевидна необходимость использования основных положений грамматики конструкций и близких к ней научных направлений. Подход «GxC» (грамматики конструкций) начал разрабатываться с 1970х годов и чрезвычайно популярен в

разных направлениях современной лингвистики: [Fillmore, Kay 1993; Fillmore 1999; Fried, Östman 2004; Goldberg 1995; Goldberg 2006; Masini 2005] и многие другие; подробную библиографию см. в http://constructiongrammar.org/.

Так что же такое «конструкция»? Кажется, стало уже традицией опираться на те свойства конструкций, которые были указаны Филмором [Fillmore et. al 1988]:

- конструкции состоят из «родительских» и «дочерних» элементов, отношения между которыми не фиксированы жестко, а могут свободно комбинироваться в предложении;
- конструкции могут определять не только синтаксические, но и лексические, семантические, прагматические параметры;
- лексические единицы могут быть включены в конструкцию;
- конструкции могут (и в некоторых случаях должны) быть идиоматичными, тогда семантика конструкции как целого будет шире семантики составляющих элементов.

Множество таким образом определяемых конструкций очень неоднородно: они будут различаться степенью и типом идиоматичности, жесткости и закрепленностью определенных лексем (классов лексем).

При широком понимании такого подхода любая синтаксическая единица является конструкцией, статус такой единицы-конструкции зависит от классификации по названным параметрам. Однако наиболее важным с точки зрения функциональности конструкции является ее положение в дихотомии парадигматика vs. синтагматика – или лексикон vs. синтаксис, инвентарные vs. конструктивные единицы (по В.Б.Касевичу [Касевич 1988]).

В предельном случае мы имеем дело с ориентацией на радикальный вариант грамматики конструкций У.Крофта (Radical Construction Grammar), отрицающий композициональность конструкций, т.е. не конструкции конструируются из элементов более низких уровней иерархии (напр., слов), а слова могут

вычленяться в результате последующих процедур обработки из целостной конструкции [Croft 2001; Croft, Cruse 2004].

В рамках парадигмы корпусных и когнитивных исследований нас интересует изучение лексико-грамматических явлений (вернее было бы даже сказать: лексических и морфолого-синтаксических явлений) при восприятии и порождении (анализе и синтезе) текста. Поэтому для нас наиболее интересным является объединение идей, заложенных в моделях грамматики конструкций и различных контекстно-ориентированных моделях (от широко известной «Контекстуальной теории значения» (Contextual Theory of Meaning) Ферса (см., напр., [Firth 1957, 1968] до современных Usage based models (см. обзор в [Barlow, Kemmer 2000]). Как известно, в процедурах обработки текста происходит максимальная опора на контекст. Причем понятие «контекст» также рассматриваться в разных смыслах. Для нас контекст предполагает широкое понимание:

•минимальный контекст, в котором реализуются лексические и морфолого-синтаксические явления;

•текстовый контекст, включающий в себя фрагменты текста вплоть до текста целиком;

•контекст, предполагающий учет текстов определенного типа (заданного функционального стиля, отобранной коллекции текстов и т.д.) (подробнее см. [Ягунова 2008].

Можно было бы добавить еще одно понимание контекста: как совокупности текстового опыта человека, а также тем самым – знание языка (на основании опыта по восприятию и порождению текстов). Такое понимание «широкого контекста» в известной степени моделируется в создании и последующем изучении Национальных корпусов.

Процедуры обработки текста носят вероятностный характер. Степень связанности конструкций, по всей видимости, зависит от вероятностной модели, описывающей ее появление в ходе процедур обработки текста; а вероятные оценки могут быть получены лишь на основании статистических данных. Причем статистические характеристики должны описывать данные в зависимости от перечисленных выше типов контекста.

Теперь обратимся к другому заявленному в названии термину. Что же из себя представляет «**коллокация**»? Сравним несколько определений этого понятия. «Collocations of a given word are statements of the habitual or customary places of that word[1]» [Firth 1957: 181]. «A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things[2]» [Manning, Schutze 1999: 141].

В отечественной 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[3]). Чаще всего принцип выделения коллокаций (в идеале список) отражает традицию определенной школы (и собственную интуицию исследователя) или узко заданную изучаемую тему. Даже в традициях русистики существует огромное количество терминологических и теоретических сложностей, что отражается в различии трактовок в словарях и грамматиках. В качестве примера позволим себе цитату из предисловия к электронному ресурсу «Словарь русской идиоматики» (это один из словарных ресурсов, создаваемых на основе Национального корпуса русского языка [Кустова 2008: 2],): «... в отечественной традиции принято различать собственно фразеологизмы (идиомы), в которых исходное значение полностью переосмысляется (*медведь на ухо наступил, ломиться в открытую дверь*), и коллокации, в которых одно

---

[1]«Коллокации заданного слова – это установления обычных или привычных мест этого слова».

[2]«Коллокация – это выражение, состоящее из двух или более слов, которое соотносится с некоторым конвенциальным способом высказывания».

[3]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].

слово выступает в своем обычном значении, а другое – во фразеологически связанном (*плакать* ***навзрыд, в стельку*** *пьяный*)». Это предисловие как бы примиряет отечественные традиции и современные парадигмы корпусной лингвистики. Все чаще приходится признавать, что, несмотря на явную неоднородность выделяемых списков, границы между классами оказываются проницаемыми. В словаре представлены «наряду с настоящими идиомами (фразеологизмами, ср. *круглый сирота*) и коллокациями (ср. *плакать навзрыд, диаметрально противоположный*), менее идиоматичные (ср. *глубоко огорчен*), а также свободные (семантически мотивированные, ср. *чрезвычайно огорчен*) сочетания со значением высокой степени)» [Кустова 2008: 2]. Такое решение создателей ресурса отвечает основным задачам контекстно-ориентированных и корпусных исследований.

Попытки последовательно учитывать контекст (причем – как указывалось выше – разные типы контекстов) ставят перед исследователем дополнительные задачи. Обычно получаемые в работах списки коллокаций лишь в некоторой степени могут быть соотносимы с исследованием тех особенностей, которые не просто заложены в языке (всех текстах на этом языке), но в существенной степени зависят от типа контекста (напр., от функционального стиля текстов, конкретной коллекции или отдельного текста по отношению к этой коллекции).

Реализовать контекстно-ориентированный подход можно с использованием различных статистических мер, позволяющих автоматически выделить из текстов коллокации и ранжировать их по степени неслучайности в соответствии со значениями выбираемых мер [Stubbs 1995]. При этом нечеткое и интуитивное понятие контекста принимает черты объективности – в узком смысле под контекстом понимается та коллекция, на которой проводится исследование. Возможность варьировать коллекции (например, выбирая коллекции текстов разных функциональных стилей или даже отдельные тексты из этих коллекций) позволяет получать списки коллокаций, различающие различные контексты. Именно текстовый материал, реализация лексико-грамматических и синтаксических проявлений, оказывается базой для исследования.

1.2. Терминологическое обобщение для решения задач данной работы

Во всех проводимых нами работах под **коллокациями** мы понимаем 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). Такой подход определяется тем, что главным для нас является опора на контекст – на коллекцию текстов или даже единичные тексты из этой коллекции. Мы в своем исследовании языка и речи идем от **реализации**, от имеющегося в нашем распоряжении материала. Именно материал диктует возможность выбора тех или иных теоретических положений и принципов классификации.

Разумеется, любое лингвистическое исследование в той или иной степени опирается на языковой материал, однако наш подход опирается на идеи **сплошного** анализа языкового материала, т.е. последовательного рассмотрений **всех** цепочек словоупотреблений определенной длины, встретившихся в исследуемой коллекции. Понятно, что такое исследование может проводиться только с использованием статистических мер, позволяющих оценивать степень неслучайности той или иной последовательности слов. Очевидно, что список, получаемый таким образом, не оказывается вполне однородным, требует дальнейшей классификации и теоретической интерпретации. Однако эти списки отражают основные особенности контекстов: различных коллекций новостных и научных текстов (и их подвыборок), а также отдельных текстов [Ягунова, Пивоварова 2010а; Ягунова, Пивоварова 2010б; Пивоварова, Ягунова 2010; Пивоварова 2010]. Идея настоящей статьи родилась в ходе анализа большого и неоднородного материала, полученного в ходе разнообразных вычислительных экспериментов. Казалось бы, существуют хорошо разработанные принципы описания, напр., разработанные И.Мельчуком основные принципы описания «фразем» (четыре стратегии, на основании последовательного применения которых можно получить 54 типа «фразем» [Mel’chuk 1995]). Однако, как уже говорилось, необходимость сплошного анализа неоднословных единиц,

выделяемых с помощью статистических мер (контекстно ориентированного анализа), потребовала иного подхода.

Понимание терминов «коллокация» и «конструкция», как уже было сказано, оказывается различным в зависимости от той или иной парадигмы. Во многих случаях одни и те же единицы могут быть названы и «коллокацией», и «конструкцией»[4]. Если пытаться разделить эти термины «по совокупности пониманий», то получится некоторое градуальное противопоставление: т.е. «скорее конструкция» vs. «скорее коллокация».

Мы предлагаем некоторую схему классификации, задающей основные параметры такого разделения. В ходе наших исследований эта схема оказалась плодотворной. Однако на настоящем этапе положения данной классификации представляются набором гипотез, которые несомненно надо верифицировать, и верификация должна происходить именно с опорой на контекст как материал анализа.

Чаще всего, термин «коллокация» используется при решении задачи выделения и описания неоднословных ***номинаций*** (не только в прикладной области). Ср. примеры из [Halliday 1966: 150]: *strong vs. powerful tea* ‘сильный vs. *сильный чай’, т.е. сочетаемостные ограничения, диктующие выбор прилагательного strong для ‘сигарет, чая и кофе’ (cigarettes, tea and coffee), но powerful, напр., для ‘героина’ (heroin). Неоднословные номинации наподобие *белый медведь*, *белый гриб*, *белое вино* или *проливной дождь*, *заклятый враг* очевидным образом ложатся в таким образом понимаемую идею коллокаций. Более того, такие традиционные признаки как «устойчивость» и «идиоматичность» (ср. [Мельчук 1960]) в известной степени переосмысляются. Колокации выходят за пределы исследования «чистой фразеологии», зачастую их целостность как единой номинации оказывается более значимым признаком, а под устойчивостью понимается скорее степень неслучайности совместной встречаемости слов. Такое понимание устойчивости

[4] Отдельно стоит прагматический признак: в прикладных исследованиях автоматической обработки текста, как правило, можно встретить термин «коллокация».

ощущается носителем языка и может быть выявлено в ходе экспериментов с информантами. Так, например, для анализируемых нами новостных и научных текстов среди таких коллокаций выступают самые разные с лингвистической точки зрения неоднословные номинации: *непосредственная близость*, *стихийное бедствие*, *Нижний Новгород*, *Саудовская Аравия*, *Бритни Спирс*, *Невский экспресс* и *корпусная лингвистика*, *речевой акт*, *именительный падеж*, *речевой сигнал*, *концептуальный граф*, *внешний посессор* соответственно.

Таким образом, коллокации достаточно часто выступают в качестве важной и частотной единицы словаря. Ср. цитату «Lexical unit is a word or collocation[5]» в начале аннотации к статье [Daudaravicius 2010]. Действительно, практические задачи автоматической обработки текста (напр., информационный и фактографический поиск) чаще всего связаны с поиском и идентификацией разнообразных сложных номинаций. Таким образом выделяются неоднословные термины, могут определяться предметные области и ключевые словосочетания, характеризующие заданную коллекцию текстов или ее подвыборку, и т. п. Именно коллокации, соответствующие неоднословным номинациям, по всей видимости могут претендовать на статус «ядерных коллокаций». В этом смысле можно было бы представить себе даже более представительную шкалу: от слова до коллокации, от колокации к конструкции. Тогда «коллокация» будет представляться как бы в виде промежуточного звена и перевалочного пункта при движении от слова к конструкции.

Конструкции, напротив, чаще всего представляют собой единицы скорее синтаксического плана. Таким образом, типовые или ядерные коллокации и конструкции часто могут оказаться противопоставленными как парадигматические vs. синтагматические единицы (или единицы, принадлежащие лексикону vs. синтаксису).

Впрочем, и здесь проявляется неоднозначность, т. к. предикативные образования, обладающие высокой степенью воспроизводимости и/или идиоматичности, будут по всей

[5] «Лексические единицы – это слова или коллокации».

видимости распределены по шкале(-ам) движения от коллокации к конструкции ближе к конструкциям. Приводимые выше *медведь на ухо наступил, ломиться в открытую дверь, плакать навзрыд, в стельку пьяный* и т.д. окажутся в зоне конструкций именно благодаря ярко выраженной предикативности. Однако для того, чтобы о них зашла речь, необходимо, чтобы они оказались реализованными в текстах и – соответственно – выделимыми с помощью статистических мер. Те, кто работает с корпусами, знает, что многие фразеологизмы в текстах встречаются довольно редко.

Отдельное внимание обратим на одно из традиционных свойств конструкций по Филмору [Fillmore et. al 1998]: лексические единицы могут быть включены в конструкцию. Следовательно, существует противопоставление с точки зрения включенности фиксированных лексем (вернее словоформ) или лексем, принадлежащих фиксированной лексико-семантической группе: напр., *А еще **N** называется!* (*А еще **друг** называется!*) (один из многочисленных примеров «синтаксических фразем», собранных и проанализированных в диссертационном сочинении М. Копотева [Копотев 2008: 125]). К данному типу конструкций примыкают разнообразные клише и устойчивые сочетания. Забегая вперед, упомянем, что такого типа конструкции – напр., «введения источника информации» – высокочастотны в новостных текстах: *сообщает РИА* 17081, *сообщает агентство* 10590, *пишет газета* 7722, *передает агентство* 7683, *передает РИА* 4487 (эта часть нашего анализа осуществлялась на коллекции [Клышинский и др. 2010], около 300 миллионов словоупотреблений). Для информационно насыщенных коллекций (наподобие lenta.ru, подробнее см. следующий пункт) конструкции, выделяемые на основании статистических мер, могут достигать длины более 5 словоупотреблений (напр., *сообщает интерфакс со ссылкой на **источник в правоохранительных органах*** из *сообщает интерфакс со ссылкой на **N***). Полагаем, что именно такой тип единиц занимает место «прототипической конструкции» на шкале(-ах) от колокации к конструкциям.

Отдельного внимания заслуживает производная служебная лексика (напр., предлоги *в течение*, *в качестве*) и дискурсивные

слова (напр., *по крайней мере*, *может быть*). Они чаще всего выступают под маркой «сочетаний, эквивалентных слову», хотя степень устойчивости этих единиц может существенно различаться, что в частности находит отражение в словарях (напр., [Богданов, Рыжова 1997]). Где они должны быть сосредоточены на шкале(-ах) движения от коллокации к конструкции? Полагаем, что в качестве условного приближения можно допустить, что они расположены в некоторой серединной зоне, равноудаленной и от «ядерных коллокаций», и от «ядерных конструкций». Это зона распределения соответствующих «сочетаний, эквивалентных слову» (термин заимствован из «Толкового словарь сочетаний, эквивалентных слову» Р.П. Рогожниковой [Рогожникова 2003], но, конечно, принципы выделения и множество единиц существенно отличается от того, что представлено в словаре). Чем выше предикативность (особенно для дискурсивных слов и наречных образований), тем они оказываются ближе к конструкциям. Другим параметром является степень устойчивости, чем выше она, тем эти единицы оказываются ближе к полюсам сосредоточения коллокаций как целостных единиц словаря (мы сейчас абстрагируемся от лингвистического анализа процессов фразеологизации).

1.3. Постановка задачи

Целью настоящей статьи является обсуждение:

- возможных направлений исследования неоднословных единиц, выделяемых статистическим образом;
- статистических характеристик, описывающих тип и степень неслучайности (устойчивости);
- выявление зависимости статистических характеристик и списков выделяемых единиц – коллокации и конструкций – от контекста, причем от контекста разных типов; в данном случае речь идет о контекстах трех видов:
    - коллекциях текстов разных функциональных стилей, и разной степени однородности,
    - подвыборках из этих коллекций,
    - отдельных текстах в рамках рассматриваемых коллекций.

При рассмотрении названных вопросов основное внимание уделяется вопросу классификации и интерпретации принципов статистического моделирования и самих выделяемых единиц в шкале «от коллокации к конструкции». Этот вопрос является основополагающим и в то же время животрепещущим и неоднозначно решаемым.

В ходе наших исследований мы выбрали в качестве базовых две статистические меры: MI и t-score[6] (подробнее о методике см. п.2.2.; [Ягунова, Пивоварова 2010а; 2010б; Пивоварова 2010]. Мы остановились на этих мерах, т.к. их специфика выделяет «коллокации» (в прикладных работах под коллокациями могут пониматься и конструкции) двух полярных типов [Manning, Schutze 2002; Stubbs 1995]. Эти полярные типы в известной степени соотносимы с полюсами на предлагаемой шкале.

Так, напр., в этих наших работах выдвигались и верифицировались гипотезы о том, что:

a. коллокации, выделяемые с помощью меры MI, чаще всего являются сложными номинациями (терминами, наименованиями объектов, ключевых для определения предметной области),
b. критерий t-score направлен, прежде всего, на выделение «устойчивых конструкций», клише и «общеязыковых устойчивых сочетаний» (производных служебных слов, дискурсивных слов) [Ягунова, Пивоварова 2010а].

Однако для решения поставленных вопросов существенную роль играет тип анализируемого контекста. Особенно ярко описанное противопоставление единиц, выявляемых на основе меры MI и меры t-score, проявляется для коллекций текстов одного функционального стиля, но неоднородных в области тематики [Ягунова, Пивоварова 2010а, 2010б].

---

[6] В своих исследованиях мы использовали также меру логарифм правдоподобия (LL), главным образом для получения единого списка, объединяющего MI-сочетания и t-score-сочетания. Однако в рамках данной статьи мы его игнорируем. Также в дальнейшем планируется использование других статистических мер.

Для монотематической коллекции текстов одного функционального стиля наблюдается зона пересечения: сочетания, выделяющиеся на основании обеих мер. Предполагаемая причина этого: формирование подможества единиц, «общих для рассматриваемой коллекции». Влияние монотематичности коллекции как особого типа контекста исследуется нами на примере материалов конференции «Корпусная лингвистика» [Пивоварова, Ягунова 2010; Ягунова, Пивоварова 2010б].

Небезынтересны случаи выделения клише и конструкций с помощью меры MI, т.е. основываясь на выраженных сочетаемостных ограничениях. Особенно ярко MI- и t-score-клише-и-конструкции противопоставлены для новостной коллекции. MI-клише и MI-конструкции носят более казенный и (квази)терминологический характер: *злоупотребление должностными полномочиями, причинение тяжкого вреда* и т. п.

Таким образом, опираясь на различный контекст (см. раздел 2) мы предлагаем в этой статье «схему движения» с указанием потенциальных нечетких границ и возможных смещений.

2. Материал и методика

2.1. Материал

В качестве основного материала использовались три коллекции текстов:

- портала www.lenta.ru с апреля по декабрь 2009; общий объем проанализированных текстов: более 66000000 «токенов» (словоупотреблением и знаков препинания);
- материалов конференции «Корпусная лингвистика» 2004-2008 года (монотематическая коллекция); объем коллекции составляет около 220000 «токенов»;
- материалов международной конференции «Диалог» «Компьютерная лингвистика и интеллектуальные технологии» за 2003-2009 годы; объем коллекции составляет около 2,5 миллионов «токенов».

Морфологическая разметка коллекции осуществлялась В.В. Бочаровым 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. При выделении коллокаций учитывались знаки препинания: рассматривались любые последовательности слов в тексте, не разделенных знаками препинания.

2.2. Методика

2.2.1. Вычислительный эксперимент с использованием мер MI и t-score. Как уже было сказано, на данном этапе нами использовались две меры *MI* и t-score (см. формулы 1 и 2 для биграмм, их обобщения для n-грамм – 1a и 2a).

$$(1)\ MI=log_2 \frac{f(n,c)\times N}{f(n)\times f(c)},$$

$$(2)\ t-score=\frac{f(n,c)-\frac{f(n)\times f(c)}{N}}{\sqrt{f(n,c)}}$$

где

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.

(1) С точки зрения теории вероятности, мера MI (mutual information, коэффициент взаимной информации) является способом проверить степень независимости 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, а значение меры MI равно нулю.

Значение меры MI зависит от размера корпуса — чем больше исследуемый корпус, тем выше в среднем получаемые по нему значения MI. Это свойство, неоднократно проверенное нами в экспериментах, по всей видимости, связанно с недостаточным объемом исследуемых нами коллекций. Теоретически, при условии «достаточно большого корпуса», где частоты слов/словосочетаний зависят только от их вероятностей, значение меры MI не должно зависеть от размера корпуса. На практике же частоты напрямую связаны с вероятностями только для отдельных слов, а для словосочетаний даже на самых больших из наших коллекций (десятки миллионов токенов) имеют место краевые эффекты. Однако создание настолько больших коллекций не всегда возможно: коллекции научных статей в принципе не очень большие; иногда в качестве «контекста» исследования выступают новостные тексты определенного издания за месяц или неделю, которых также не очень много (в качестве примера такого исследования см. [Пивоварова 2010]).

Зависимость меры MI от размера корпуса затрудняет сравнение значений мер, полученных на разных корпусах, или например, на полной коллекции и ее части. Один из способов решения это проблемы, используемый нами в данной работе, это игнорирование числового значения меры MI и использование ее только в качестве средства 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, окказионализмами, иностранными словами и другим информационным шумом, который неизбежен в большой коллекции. Поэтому для данной меры используется порог

отсечения по частоте. К сожалению, правильный подбор порога отсечения оказывается чрезвычайно сложной задачей:

- при его определении исследователь чаще всего опирается на задачу исследования, в рамках которой он определяет требуемые пределы точности и/или полноты выборки;
- определяются основные характеристики коллекции – не только объем, но и степень однородности и монотематичности;
- далее необходимо проводить отдельный вычислительный эксперимент, чаще всего ограничивающимся подбором значений порогов с последующим экспресс-анализом получаемых выборок и распределений значений мер.

Таким образом, для каждой коллекции подбирается свое пороговое значение.

В данной работе мы рассматривали только те биграммы, которые встретились в коллекции не менее сорока раз для коллекции «лента.ру» и «Диалог» и не менее шестнадцати раз для корпусной коллекции. Высокие пороги отсечения определялись самой постановкой задачи: нашей целью является выделить наиболее значимые, характерные для данной коллекции коллокации и конструкции, т.е. акцент делается на точности, а не на полноте.

(2). 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.

Данная мера используется гораздо реже, чем мера MI, в частности, потому что 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 [Stubbs 1995] 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. В литературе эта особенность часто трактуется как малопригодность этой меры для поиска терминологических словосочетаний и номинаций; для этой цели она, как правило, не используется. Естественно, что мера t-score, в отличие от MI, не преувеличивает значимость редких коллокаций и не требует использования 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.

В нашем исследовании мы учитывали порядок коллокатов внутри биграммы.

Меру MI можно обобщить для любого числа коллокатов (на эту тему см. напр., [Petrovic et al. 2006]). Нами проводились исследования коллокаций, включающих от двух до пяти коллокатов. Различные варианты обобщения этой меры (напр., [Boulis 2002; Petrovic et al. 2006; Tadic, Sojat 2003; Su, Wu, Chang 1994]) представляют тему отдельного исследования. В данной статье мы рассматриваем результаты, полученные с помощью следующего варианта:.

$$\text{(1a)}\quad MI = \log_2 \frac{f(n, c_1, c_2) * (N^{**}(i-1))}{f(n) * f(c_1) * f(c_2)},$$

где i – число коллокатов, остальные условные обозначения те же, что и для формул 1 и 2.

Обобщение меры t-score для коллокаций длинне, чем биграммы, в литературе практически не встречается. Причиной этого может быть тот факт, что мера t-score является аппроксимацией частоты, которая за счет поправочного коэффициента «понижает» значимость словосочетаний, состоящих из двух очень частотных слов (например, двух союзов или союза и предлога). Поскольку сами коллокаты очень частотны, такие коллокации становится частотными просто в силу вероятностных причин. Однако чем больше число коллокатов входит в коллокацию, тем меньше сила этого эффекта (не говоря уже о сомнительности появления в тексте, например,

трех союзов подряд). Поэтому для многословных коллокаций использование t-score не представляется осмысленным, а сама частота становится более надежным источником информации, чем для биграмм. В нашей работе для многословных сочетаний используется собственно частота коллокации (вместо расширенного варианта t-score).

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, 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.

Методика первичной обработки выдач (списков сочетаний), полученных на разных коллекциях и с помощью разных мер включала удаление из первоначальных списков тех сочетаний, которые включали слово(-а), написанные латиницей. Затем биграммы упорядочивались по убыванию значения меры MI или t-score. Главное внимание при классификации и интерпретации уделялось первым 100 биграммам из получившихся списков.

2.2.2. Серия экспериментов по оценке совместной встречаемости и взаимного притяжения слов с опорой на ближайший контекст. Кратко, насколько позволяет формат данной статьи, остановимся на проекте, реализуемом в настоящее время. Об окончательных результатах этого проекта говорить еще рано, несмотря на это упоминание проекта и используемой в нем методики, на наш взгляд более чем логично в рамках данной статьи. Именно в нем реализуется подход, позволяющий сопоставлять результаты, полученные на материале коллекции и на материале отдельного текста из этой коллекции. Тем самым реализуется возможность наиболее точного учета контекста. Более того, этот подход кажется нам изначально ориентированным на соединение антропоцентричности и вычислительных процедур в рамках исследования обработки текста.

Этот проект предполагает сочетание вычислительного эксперимента и эксперимента с информантами. В ходе вычислительного эксперимента меры совместной встречаемости высчитывается на основании коэффициента Дайса (Dice, см. (3)).

$$(3)\quad Dice(x,y) = \frac{2 * f(x,y)}{f(x) + f(y)},$$

где f(x) и f(y) – частота встречаемости слов x и y в коллекции, а f(x,y) – частота совместной встречаемости слов x и y.

Практически более удобным оказывается использовать видоизмененную меру Дайса (см. 3а). Эта мера оказывается сходной с широко используемой мерой MI, но авторы метода находят ее более применимой для дальнейшей разметки коллекции и отдельных текстов этой коллекции [Daudaravicius 2010].

$$(3a)\quad Dice'(x,y) = \log_2\left(\frac{2 * f(x,y)}{f(x) + f(y)}\right),$$

Процесс вычислительного эксперимента можно коротко описать следующим алгоритмом. Сначала для всех пар слов по всей коллекции считается коэффициент Дайса. Затем для каждого конкретного текста, представляющего собой цепочку слов или вернее цепочку пересекающихся пар (слово *x* с предшествующим словом и слово *x* с последующим словом), осуществляется «сборка» связанных сегментов. При последовательном прохождении от слова к слову в каждом тексте уже известны соответствующие значения коэффициента Дайса для всех пересекающихся пар. На основании значений этой статистической меры слова объединяются в связанные группы с учетом ближайшего контекста (принимается решение о том, надо ли присоединить текущее слово к предыдущему). Слово не присоединяется к предыдущему, если значение коэффициента Дайса для данной пары ниже порогового, или если оно ниже, чем среднее арифметическое того же коэффициента для левой и правой пары. Во всех остальных случаях слово присоединяется.

Текст в итоге выглядит следующим образом: A_B_C D_F. То, что связано знаком подчеркивания – воспринимается программой как связанный сегмент текста (коллокация или конструкция), там, где такого знака нет, проходит граница между сегментами. Сегмент может включать произвольное число слов.

В результате такого вычислительного эксперимента мы получаем два набора: набор связанных биграмм по коллекции (упорядоченный по убыванию значения меры) и набор связанных сочетаний, подсчитанных для каждого текста отдельно, а затем объединенный в некое подобие частотного словаря связанных сочетаний. Программа, реализующая этот алгоритм, доступна для скачивания с сайта ее создателя: http://donelaitis.vdu.lt/~vidas/tools.htm.

Нас интересует именно результат сравнения связанности в корпусе (по мере Дайса) и связанности в рамках контекста, заложенного в конкретном тексте. Более того, результаты данного вычислительного эксперимента, как мы полагаем, должны в значительной степени соотноситься с процедурами анализа текста испытуемыми.

Носитель языка имеет интуитивные представления о неслучайно встречающихся сочетаниях слов: текстовые базы по текстам разных функциональных стилей. На основании этого знания адресат воспринимает каждый конкретный текст, не противоречащей его текстовой базе. Проводимый в настоящее время эксперимент с информантами представляет собой оценку связности между (пробельными) словами в шкале от 0 до 5, где 5 – соответствует максимальной, а 0 – минимальной степени связности. Получив эти данные от информантов, мы сможем выстраивать более длинные цепочки слов по алгоритму, аналогичному описанному выше, и затем сравнивать сегментацию текста, произведенную автоматически, с сегментацией, полученной от информантов.

3. Результаты

3.1. MI-коллокации

Как уже говорилось, под типичными коллокациями в нашей классификации мы понимаем прежде всего неоднословные номинации и сложные термины. Более того, такие колокации

зачастую выходят за пределы «чистой фразеологии», их целостность как единой номинации оказывается более значимым признаком, а под устойчивостью понимается скорее степень неслучайности совместной встречаемости слов.

Коллокации достаточно часто выступают в качестве важной и частотной единицы словаря. В этом смысле «ядерные» колокации могут рассматриваться не только на шкале от «коллокации до конструкции», но и на дополнительной шкале «от слова до коллокации».

А что такое «слово»? Не углубляясь в неоднозначность определения – казалось бы – ведущей единицы языка и речи, вспомним о наличии противоречий даже на этом уровне. Что является единицей анализа текста: лексема или словоформа? Можно считать более чем обоснованным и экспериментально доказанным положение о том, что словоформа является ведущей единицей анализа текста (лексема выполняет роль дополнительной единицы анализа, востребуемой лишь в особых случаях) [Касевич, Ягунова 2004; Касевич, Ягунова 2006].

При работе с коллокациями выбор основной единицы анализа представляет собой дополнительный вопрос.

Разберем возможности решения этого вопроса на примере биграмм: полагаем весьма показательным сопоставление биграмм, выявленных для лексем и/или для словоформ[7].

На материале новостных текстов был проведен предварительный сопоставительный анализ (1) списка сочетаний, выделяемых для лексем (но не словоформ), (2) списка сочетаний, выделяемых для словоформ (но не лексем) и (3) списка сочетаний, выделяемых и для лексем, и для словоформ (подробнее см. статью [Ягунова, Пивоварова 2010а])[8].

---

[7]Хочется отметить, что различные аудитории, обсуждавшие доклады на эту тему высказывались весьма «категорично»: некоторые аудитории лишь лексемные коллокации считали достойными внимания, другие – напротив – только словоформные. Безусловно, основные особенности, рассмотренные на примере биграмм-коллокаций, действуют и при увеличении объема сочетания.

[8]Во всех трех случаях под «списком» имеется в виду первая сотня словосочетаний, выявленных тем или иным способом. Очевидно, что списки, взятые целиком, будут совершенно идентичны, т.к.

1. 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), что обозначается с помощью прописных букв):

- для новостных текстов – 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, 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;*
- для научных текстов – 9 *ВИНИТЕЛЬНЫЙ ПАДЕЖ*, 17 *ИМЕНИТЕЛЬНЫЙ ПАДЕЖ*, 24 *АКТУАЛЬНЫЙ ЧЛЕНЕНИЕ*, 29 *ИНСТРУМЕНТАЛЬНЫЙ СРЕДА*.

Наибольшую сложность представляет список на материле новостных текстов. Среди первых 100 новостных 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 (приблизительно поровну: 13 и 12 соответственно). Деление на термины и нетермины для новостных текстов довольно условно,

---

статистические меры подсчитывались для всех словосочетаний с частотой более заданной. Нас интересует, однако, словосочетания с наибольшим значением меры, т.е. верхние части списков, которые мы в дальнейшем для краткости именуем просто списками.

т.к. многие номинации, исходно носящие терминологический характер, давно и прочно вошли в общеязыковую практику[9].

Как уже было сказано, показательна высокая процентная доля, которую имеют в этом классе наименования лиц для новостных текстов (по-видимому, для многих наименований лиц особо характерна высокая степень разнообразия набора семантико-синтаксических ролей, в которых они выступают) и терминов для научных текстов. 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. Такие номинации, условно говоря, можно сопоставить со словом, которое характеризуется достаточно полной парадигмой формоизменения.

2. Биграммы второго типа, как правило, относятся к номинации в определенной синтаксической позиции. Примеры биграмм этого списка (число обозначает п.п.):

- для новостных текстов – 3 *парниковых газов*, 5 *Соединенных Штатов*, 6 *Женской Теннисной*, 10 *кредитном портфеле*, 11 *Палестинской автономии*, 13 *встречную полосу*, *Нижнем Новгороде*, 18 *Федеральную трассу*;
- для научных текстов – 10 *речевой акт*, 50 *речевых актов*, 19 *именная группа*, 65 *именных групп*, 27 *коммуникативного акта*, 62 *коммуникативных актов*, 77 *просодических характеристик*, 78 *прошедшего времени*, 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*».

Среди первых 100 биграмм из словоформ встретились повторения лишь трех номинаций: *Саудовская Аравия* и

---

[9] Для тех новостных 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).

Саудовской Аравии, *Бараком Обамой* и *Бараку Обаме*, печально известного названия ночного клуба *Хромой лошади* и *Хромая лошадь*. 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 выше доля сочетаний, претендующих на устойчивость в качестве сложных номинаций и терминов, чем для лексемных биграмм[10].

В этих списках в обоих случаях некоторая составная номинация или термин 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 типичной для них семантико-синтаксической функции. Такое тяготение оказывается важным и для однословных номинаций и для неоднословных[11].

В том случае, если данная составная номинация входит в состав некоторого более крупного – трёхсловного или даже более протяжённого, напр., *(Женской теннисной) ассоциации*, *теории (речевых актов)* – сочетание является более устойчивым на синтагматической оси, чем в случае прочих словоформных биграмм (допускающих более свободные связи с соседями на синтагматической оси).

---

[10]Словоформы как единицы биграмм демонстрируют морфологически оформленные синтаксические отношения. В анализируемой части этих новостных 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).

[11] О действии такого закона применительно к словам писали, в частности, Н.Д.Арутюнова, Г.А.Золотова, В.Г.Гак, Ю.Д.Апресян.

3. Биграммы третьего класса занимают в текущем словарном составе некое **промежуточное место** между биграммами класса «1» и биграммами класса «2». Это 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[12].

На наш взгляд (т.е. в результате пристального изучения списков), этот третий класс – биграммы, выделяющиеся и для лексем, и для словоформ – оказывается, как правило, наиболее информативным. Таким способом мы выделяем наиболее информационно-нагруженные и точные сочетания, характеризующие данную коллекцию (см. напр., биграммы в Таблицах 1, 2 и 3). Для простоты восприятия в таблицах биграммы представлены в виде сочетаний словоформ (соответствующей словоформной биграмме). Ведущее место в нем отводится интересующим нас «ядерным коллокациям». Однако в таблице присутствуют и сочетания, рассматриваемые нами в следующем пункте **MI-конструкции** (особенно для научных коллекций).

[12] Причина попадания в класс «3» может быть и в отсутствии формальной морфологической оформленности: в него 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.

Таблица 1. Пример пересечения между биграммами для лексем и для словоформ (для первой сотни)[13]

пп. (для лексем)	пп. (для словоформ)	биграммы
1	1	Бритни Спирс
2	2	Эльвира Набиуллина
3	23	Ле Бурже
9	36	Лионель Месси
10	4	мысе Канаверал
11	43	бин Ладена
14	9	Норильского никеля
15	7	дельты Нигера
17	50	Ак Барс
18	28	тротиловом эквиваленте
19	20	тройскую унцию
20	70	Ролан Гаррос
26	49	дель Торо
27	87	дель Потро
29	33	Арбат Престиж
31	96	РАО ЕЭС
32	35	Салават Юлаев
34	51	Арсений Яценюк
36	42	голубых фишек
39	29	адронного Коллайдера

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

Таблица 2. Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Корпусная лингвистика»[14]

п.п.	Биграммы	п.п.	Биграммы
2	наш взгляд	36	одной стороны
3	(по) крайней мере	37	таким образом
4	речевой деятельности	40	разрешения неоднозначности
5	художественной литературы	41	английский язык
7	первую очередь	43	кроме того
9	общим объемом	47	Национальный корпус
11	корпусная лингвистика	48	грамматических категорий
13	имена собственные	52	устная речь
15	математической лингвистики	54	база данных
16	словарной статьи	58	во многих
17	свою очередь	61	лексических единиц
18	предметной области	62	дает возможность
19	машинного перевода	63	зависит от
20	точки зрения	64	отличие от
22	за счет	65	русский язык
24	речь идет	67	корпусные данные
25	прежде всего	68	отличается от
26	большое количество	71	зависимости от
28	настоящее время	72	работы над
31	представляет собой	79	частей речи
32	млн словоупотреблений	80	во всех
34	другой стороны	84	при помощи
35	семантических состояний	86	морфологической разметки

[14] Большую длину списка мы связываем с большей однородностью данной коллекции.

Таблица 3. Биграммы (MI-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Диалог».

п.п.	Биграммы	п.п.	Биграммы
1	ударном слоге	28	интеллектуальные технологии
2	концептуальных графов	30	корпусная лингвистика
4	внешним посессором	33	отглагольных существительных
5	оперативной памяти	37	знаки препинания
8	вокального жеста	38	педагогической коммуникации
14	крайней мере	42	основного тона
16	XIX века	46	машинного перевода
17	лингвистического процессора	61	устойчивых словосочетаний
21	положение дел	63	точки зрения
22	первую очередь	70	меньшей мере
25	картине мира	72	вряд ли
26	множественного числа	73	предметной области
		85	вплоть до

(по) крайней мере, (в) первую очередь, (с) точки зрения, (по) меньшей мере, прежде всего

3.2. MI-конструкции

Большинство клише и конструкций выделяется с помощью меры t-score. Однако некоторые типы клише и конструкций хорошо извлекаются с помощью меры MI (т.е. основываясь на выраженных сочетаемостных ограничениях). Особенно эти разные типы противопоставлены для новостной коллекции. Прежде всего, эти MI-клише и MI-конструкции носят более казенный и (квази)терминологический характер: *злоупотребление должностными полномочиями, причинение тяжкого вреда* и т.д.

Если для новостных биграмм отмечены лишь штучные варианты: конструкция *НАЧИНИТЬ ВЗРЫВЧАТКА* для лексем и *обогащению урана* для словоформ, то в списках триграмм для новостной коллекции клише и конструкции составляют более 30% (30% для словоформ и 35% для лексем).

Примеры:

для лексем – *УМЫСЛИТЬ ПРИЧИНЕНИЕ ТЯЖКИЙ, КРАТКИЙ ИЗЛОЖЕНИЕ ПРИВОДИТЬСЯ, ПОДРЫВ НЕВСКИЙ ЭКСПРЕСС, ПРЕВЫШЕНИЕ ДОЛЖНОСТНОЙ ПОЛНОМОЧИЕ, ПСИХОЛОГИЧЕСКИ ВАЖНЫЙ ОТМЕТКА, ДА ПРИЙТИ СПАСИТЕЛЬ, ТЯЖКИЙ ВРЕД ЗДОРОВЬЕ, ВРЕМЕННО НЕДЕЙСТВУЮЩИЙ ЧЕМПИОН, ЗАСЛУГА ПЕРЕД ОТЕЧЕСТВО, ЭКОНОМИЧЕСКИ АКТИВНЫЙ НАСЕЛЕНИЕ, КРАТКИЙ ИЗЛОЖЕНИЕ ПРИВОДИТЬ, ЗЛОУПОТРЕБЛЕНИЕ ДОЛЖНОСТНОЙ ПОЛНОМОЧИЕ, СОСТОЯНИЕ АЛКОГОЛЬНЫЙ ОПЬЯНЕНИЕ, НАПИСАНИЕ ДАННЫЙ ЗАМЕТКА, ДАТЬ ПРИЗНАТЕЛЬНЫЙ ПОКАЗАНИЕ, ПАДЕНИЕ БЕРЛИНСКИЙ СТЕНА, КРУШЕНИЕ НЕВСКИЙ ЭКСПРЕСС, ОБЪЕДИНИТЬ АВИАСТРОИТЕЛЬНЫЙ КОРПОРАЦИЯ, РАЗЛИЧНЫЙ СТЕПЕНЬ ТЯЖЕСТЬ, ПОКОНЧИТЬ ЖИЗНЬ САМОУБИЙСТВО, ОСВОБОДИТЬ ИЗ-ПОД СТРАЖ;*

для словоформ – *злоупотреблении должностными полномочиями, причинение тяжкого вреда, написания данной заметки, превышении должностных полномочий, краткое изложение приводится, совершил аварийную посадку, покончил жизнь самоубийством, превышение должностных полномочий, произошла массовая драка, сработало взрывное устройство, краткое изложение приводит, числятся пропавшими без, такому выводу пришли, фондовые индексы завершили, выглядит следующим образом*.

Более того, приведенные примеры иллюстрируют то, что многие из конструкций имеют явно выраженную предикативность. Так, конструкции с глаголом в вершине составляют 12% от MI-конструкций для словоформ, и 11% – для лексем; кроме того, еще с отглагольным существительным – в 6% случаев для словоформ и в 7% для лексем.

Граница между клише и конструкциями во многих случаях нечеткая. Так, напр., *должностные полномочия* могут сочетаться с *злоупотреблением* или *превышением*, с *злоупотреблять* или *превышать*. Общая логика заставляет предполагать чуть большую близость к конструкциям в случаях с глагольной вершиной. По-видимому, можно выделить два фактора, в какой-то степени разводящих клише и конструкции: глагольность и

интуитивно ощущаемый казенно-канцелярский аромат сочетаний. Наиболее «правильными» среди выделяемых сочетаний полагаем конструкции типа *такому выводу пришли, фондовые индексы завершили, выглядит следующим образом.*

Граница между коллокациями и клише также нечеткая. Результаты анализа полученных списков позволяют предполагать, что признаками, которые можно считать условно разделяющими коллокации и клише, являются казенный колорит и референциальный статус. Под последним признаком мы понимаем то, что «коллокации» чаще всего включают в себя сложные номинации, обозначающие уникальный объект (или чрезвычайно информационно важный – для рассматриваемого контекста, напр., коллекции – класс объектов) внеязыковой действительности, коллокации-«клише», как правило, относятся к «традиционным» и сравнительно большим классам объектов внеязыковой действительности, напр., коллокациями-клише будут *ВЕТЕРАН ВЕЛИКИЙ ОТЕЧЕСТВЕННЫЙ, КОЛОНИЯ СТРОГИЙ РЕЖИМ, САМОДЕЛЬНЫЙ ВЗРЫВНОЙ УСТРОЙСТВО*.

В целом, можно рассматривать термин «клише» как перпендикулярный к шкале «коллокация-конструкция» - он отражает скорее стилистические характеристики, а с морфосинтаксической точки зрения, как ясно из вышеприведенного обсуждения, клише может являться как коллокацией, так и конструкцией. Отметим также, что клише являются неотъемлемой частью газетного стиля, их обилие в новостных текстах, как нам кажется, можно объяснить следующими условиями:

- большое количество информации, полученной из официальных источников, и как следствие, сильное влияние официально-делового функционального стиля;
- требование оперативности, высокая скорость порождения текстов, которая приводит к многократному использованию одних и тех же шаблонов;
- высокие требования к скорости и качеству усвоения информации, которая для этого должна быть представлена в узнаваемой, всегда одной и той же форме.

Все эти условия приводят к известной шаблонности новостных текстов, существенно облегчающей их обработку в системах автоматического анализа, которые довольно плохо справляются с художественными и художественно-публицистическмими текстами.

3.3. t-score-конструкции

Биграммы, выделяемые с помощью меры t-score кажутся сравнительно легко интерпретируемыми. Даже для новостной коллекции в 80% случаев мы наблюдаем пересечение списка словоформных и лексемных биграмм (ср. табл. 4).

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)*, места и времени (*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*[15]). 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*[16] (ср. с Таблицей 4).

---

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

[16] Это, очевидно, составные части более длинных выражений «*как сообщает корреспондент*», «*по данным агенства*», «*со ссылкой на*», которые оказываются среди наиболее частотных три- и более грамм

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

Выделим несколько основных типов такого рода сочетаний для новостных текстов, маркирующих особенности новостных текстов (см. табл. 4):

- составные служебные и дискурсивные слова, напр., *в течение, в качестве, в ходе, в частности, в результате, пока не, кроме того;*
- сложные номинации, прежде всего, относящиеся к наименованиям источников информации (материал для раздела 3.4, напр., *РИА Новости*), при переходе к более объемным сочетаниям (три- и более граммам) они входят в состав конструкций «введения источника информации»;

- колокации-клише (напр., *миллионов долларов*, *миллиарда долларов*), которые при переходе к более объемным сочетаниям могут входить в состав конструкций;
- сочетания, имеющие все показатели конструкций (как правило, компоненты конструкций «введения источника информации»):
    - с глаголом – напр., *сообщает РИА*, *как сообщает*, *это сообщать*,
    - с существительным – напр., *со ссылкой*, *по ссылкам*.

Для научных текстов также выделяется ряд типов t-score-сочетаний, маркирующих научный функциональный стиль (см. табл. 2 и 3):

- составные служебные и дискурсивные слова, напр., *(по) крайней мере, (в) первую очередь, (с) точки зрения, (по) меньшей мере, прежде всего;*
- конструкции и сходные с ними составные обороты: *дает возможность, зависит от vs. (в) зависимости от, (в) отличие от vs. отличается от* и т.д.

Во введении мы сформулировали – в качестве условного приближения – предположение о том, что производная служебная лексика (напр., предлоги *в течение*, *в качестве*) и дискурсивные слова (напр., *по крайней мере*, *может быть*) расположена в некоторой серединной зоне, равноудаленной и от «ядерных кодллокаций», и от «ядерных конструкций». Чем выше предикативность (особенно для дискурсивных слов и наречных образований), тем они оказываются ближе к конструкциям. Другим параметром является степень устойчивости, чем выше она, тем эти единицы оказываются ближе к полюсам сосредоточения коллокаций как целостных единиц словаря (мы сейчас абстрагируемся от лингвистического анализа процессов фразеологизации).

Соответственно в предлагаемой схеме – в соответствии с признаком предикативности – *в зависимости от* и *в отличие от* находится ближе к середине, а *зависит от* и *отличается от* – чуть ближе к конструкциям.

Степень устойчивости определяется нами, прежде всего, на основании эксперимента с информантами (и дальнейшей лингвистической интерпретации полученных результатов). Нами

проводится оценка степени связанности сочетаний, выделенных на основании рассматриваемых двух мер, при помощи серий экспериментов с информантами[17]. Анной Савиной была проведена первая серия экспериментов, в которых испытуемые должны были оценивать целостность-связанность-неслучайность (относя сочетания к одному из трех заданных классов) приводимые в анкете сочетания, имеющие наиболее высокие значения мер. Естественно, нас интересовало влияние функционального стиля, т.ч. (1) в каждой из анкет были собраны сочетания, относящиеся только к одной коллекции, (2) информаторам сообщалось, из каких тексов - научных или из новостных – были извлечены сочетания.

Например, для научных текстов *в частности* и *с помощью* характеризуются большей целостностью и связностью, чем *в качестве*, *за счет*, *на основе*; *с одной стороны*, *с другой стороны*, *по сравнению с* и *в отличие от* характеризуются меньшей целостностью, чем *с точки зрения* и *в соответствии с*. Т.е. напр., морфологическая цельнооформленность *в отличи**е** от* не явилось для наивных носителей языка (участников этого эксперимента) решающим признаком для признания высокого уровня целостности и связности.

Аналогично, для новостной коллекции, напр., *этом сообщает*, *в результате* являются менее целостными, чем *как сообщает*, *по данным*; *сообщает РИА Новости*, *об этом сообщается* обладают большей целостностью и связностью, чем *новости со ссылкой*, *по его словам*, *об этом сообщает*. Конечно, окончательный результат будет получен на основании серии взаимодополняющих экспериментов (как по методике, так и по материалу, представленному в анкетах для испытуемых).

На рассматриваемом нами материале типичными представителями конструкций («ядерными конструкциями») являются «конструкции ввода информации» в новостных текстах. В таблице 5 мы привели верхушку списка частотных «пятиграмм» (из рассматриваемого набора только два сочетания

[17] Надеемся, что в ближайших публикациях мы сможем показать специфику принятия решения испытуемыми при оценке степени устойчивости-связности и дать лингвистическую интерпретацию основных параметров, влияющих на принятие решения.

не относились к введению источника информации; кроме того, мы не стали исключать слова, написанные латиницей, для иллюстрации того, что в состав этих конструкций в принципе могут входить наименования информационных агентств любого типа). Напомним, что пятиграммы выделялись на основании частоты встречаемости коллокации: для больших n мера t-score как аппроксимация частоты оказывается избыточной (см. раздел 2.2.1).

Таблица 5. Наиболее частотные «пятиграммы», являющиеся «конструкциями ввода информации» в новостных текстах (на материале портала lenta.ru).

РИА Новости **со ссылкой** на
сообщает РИА Новости **со ссылкой**
сообщает Интерфакс **со ссылкой** на
Со ссылкой на источник в
Об этом **сообщает** РИА Новости
(об) этом **сообщает** РИА Новости **со**
На источник в правоохранительных органах
(со) ссылкой на источник в правоохранительных
Об этом **сообщает** официальный сайт
Об этом **сообщается** в пресс-релизе
агентство Интерфакс **со ссылкой** на
Об этом **сообщает** Интерфакс **со**
(об) этом **сообщает** Интерфакс **со ссылкой**
сообщает AFP **со ссылкой** на
Об этом **пишет** газета Коммерсант
Новости **со ссылкой** на источник
Об этом **пишет** газета Ведомости
Интерфакс **со ссылкой** на источник
сообщает ИТАР-ТАСС **со ссылкой** на
сообщает агентство Интерфакс **со ссылкой**
Об этом **сообщает** Associated Press
Об этом **сообщается** на сайте
Интерфакс **со ссылкой** на пресс-службу
Об этом **говорится** в официальном
газета Ведомости **со ссылкой** на
Новости **со ссылкой** на пресс-службу

газета Коммерсант **со ссылкой** на

Наиболее частотная схема такой конструкции сводится к:

1 (*об этом*) + 2 глагол (*сообщает*, *сообщается*, *пишет*, *говорится* и др.) + 3 название информационного агентства + 4 *со ссылкой (на)* + 5 источник информации.

На материале текстов портала «лента.ру» наиболее часто в состав конструкции входит глагол *сообщает* или *сообщается*, однако это предпочтение может носить стилистический характер, отличающий именно портал «лента.ру».

На настоящее время ведется работа по исследованию конструкции «введения информации» на материале других новостных коллекций: «Компьюлента» и «Независимая газета». Уже сейчас можно говорить о том, что способы заполнения «конструкции введения источника информации» зависит от коллекции (т.е. информационного источника). Например, для «Независимой газеты» биграмма ссылкой на стоит на 1551 месте, среди словоформных биграмм, упорядоченных по значению меры t-score, а со ссылкой – на 1591-м месте. Среди лексем первая биграмма со словом «сообщать» *КАК СООБЩАТЬ* стоит на 967 месте, следующая – *СООБЩАТЬ ИНТЕРФАКС* – на 5096 и т.д. Ср. также с данными «Статистического словаря русской газеты» А.Я. Шайкевича [Шайкевич и др., 2008] *сообщается* 492, *сообщать* – 1614, *сообщаться* – 29, *сообщение* – 2488, *сообщить* – 8248 (корпус 1997-го года, 15 млн. словоупотреблений).

3.4. t-score-коллокации

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 информационно важных номинаций и терминологических словосочетаний, не используясь для этой цели.

Однако все зависит от контекста, в данном случае от степени монотематичности и однородности коллекции. Так, в процессе данной 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 образом был получен список терминологических биграмм, общих для **всех** (или **подавляющего большинства**) текстов рассматриваемых коллекций (см. Таблицы 5 и 6).

Таблица 5. Терминологические биграммы (t-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Диалог»

лексемные биграммы	**словоформные биграммы**
РУССКИЙ ЯЗЫК	русского языка
	русском языке
ПРЕДМЕТНЫЙ ОБЛАСТЬ	предметной области

Таблица 6. Терминологические биграммы (t-score), выделяющиеся и для лексем, и для словоформ. Материал конференции «Корпусная лингвистика»

лексемные биграммы	**словоформные биграммы**
РУССКИЙ ЯЗЫК	русского языка
	русский язык
КОРПУС ТЕКСТ	корпус текстов
	корпуса текстов
НАЦИОНАЛЬНЫЙ КОРПУС	национального корпуса
	национальный корпус
ЧАСТЬ РЕЧЬ	части речи
	частей речи
АНГЛИЙСКИЙ ЯЗЫК	английского языка
КОРПУС РУССКИЙ	корпус русского
	корпуса русского
МАШИННЫЙ ПЕРЕВОД	машинного перевода
СЕМАНТИЧЕСКИЙ РАЗМЕТКА	семантической разметки
ПРЕДМЕТНЫЙ ОБЛАСТЬ	предметной области
ЛЕКСИЧЕСКИЙ ЕДИНИЦА	лексических единиц
ПАРАЛЛЕЛЬНЫЙ ТЕКСТ	параллельных текстов

Сопоставление списков терминологических биграмм, общих для **всех** (или **подавляющего большинства**) текстов (t-score-биграмм-коллокаций) рассматриваемых коллекций, 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 в текстах коллекции уделяется внимание *предметной области*.

2. Представляемые на «Корпусной конференции» исследования чаще всего ориентированы на *русский язык* или *английский язык*. В качестве материала (и/или объекта исследования) в большинстве работ выступает *корпус текстов*, что *лексическим единицам* (*частям речи, семантической разметке* лексических единиц) уделяется особое внимание. Что многие исследования ориентированы на решение вопросов *машинного перевода* и связаны с текстами заранее заданной *предметной области.* Таким образом, наши выводы согласуются с традиционной тематикой корпусных исследований, что отражено в наборе «общих» терминологических сочетаний.

Причем именно биграммы (а не триграммы и далее n-граммы) дают на нашем материале наиболее информационно насыщенную картину. Впрочем, возможно, что одна из причин этого лежит в сравнительно небольшом корпусе материалов конференции «Корпусная лингвистика (см. раздел 2.1).

По-видимому, чем выше однородность коллекции, тем более информативным окажется набор подобных t-score-биграмм-коллокаций для описания коллекции как целостного информационного потока (о понятии информационного потока см. в [Антонов, Ягунова 2010]).

Заключение

В данной статье мы постарались обсудить

- возможных направлений исследования неоднословных единиц, выделяемых статистическим образом;
- статистических характеристик, описывающих тип и степень неслучайности (устойчивости);

- выявление зависимости статистических характеристик и списков выделяемых единиц – коллокации и конструкций – от контекста, причем от контекста разных типов.

Исследование контекста, который мы условно назвали «отдельные тексты в рамках рассматриваемых коллекций», на настоящее время находится на начальном этапе. В этой статье мы еще не приводим полученных результатов на основании этой нетривиальной методики. Однако саму постановку проблемы и краткое описание методики приводим, т.к. без них наш исследовательский подход смотрелся бы неполно. Эта часть – исследование такого типа контекста – позволяет проследить формирование и функционирование коллокаций при обработке текста, учитывающей ближайший текстовый контекст анализируемой единицы (ориентирующиеся на процедуры, используемые человеком при анализе текста), недаром именно в этой части максимально спаянными оказались вычислительные эксперименты и эксперименты с носителями языка.

При рассмотрении названных вопросов основное внимание уделяется вопросу классификации и интерпретации принципов статистического моделирования и самих выделяемых единиц в шкале «от коллокации к конструкции». Этот вопрос является основополагающим и неоднозначно решаемым.

Решаемые вопросы оказались краеугольными в ходе анализа большого и неоднородного материала, полученного в ходе разнообразных вычислительных экспериментов. Их диктует необходимость **сплошного** анализа неоднословных единиц, выделяемых с помощью статистических мер.

Мы предлагаем некоторую схему классификации, задающей основные параметры движения по шкале «от коллокации к конструкции» с нечеткими границами явно выраженной динамической природы. Положения данной классификации представляются набором гипотез, с одной стороны, уже верифицированных, а с другой – требующих дальнейшей верификации с учетом все большего числа параметров (прежде всего, контекстно-ориентированных параметров).

И совсем в заключение хотела бы присоединиться к тем положениям, которые сформулировала Марина Русаков в

качестве заключения своей докторской диссертации. Позволим себе большую цитату как знак согласия, дань уважения, любви и признательности: «Проведенное исследование показало, что создание описания русского (и любого другого) языка с опорой на закономерности реализации в процессе речевой коммуникации различных грамматических явлений (и лексических. – *Е. Я.* и *Л. П.*) не только необходимо, но и возможно.

Возможности создания такого рода грамматик (и словарей. – *Е. Я.* и *Л. П.*) определяются следующими факторами, характеризующими современное состояние лингвистической науки:

1. Достаточной развитостью полученных на основании изучения текстов знаний о грамматических системах конкретных языков и грамматики в целом.

2. «Воссоединением» лингвистики с другими науками о человеке.

3. Революционными процессами в области развития технологии лингвистического исследования. Для проведения исследований, сходных с настоящей работой по целям и задачам, самым важным в области методологии является возможность обращения к различным Интернет-ресурсам, в первую очередь к большим массивам данных, организованным в корпусы. Появление новых методологических возможностей позволяет решать вопросы, которые накапливались в лингвистике в течение десятилетий.

Проведенное исследование показало, что обращение к говорящему и слушающему человеку, к различным типам естественной речи позволяет лингвисту выйти за жесткие рамки идеологически окрашенных лингвистических школ и концепций и получать факты, которые не нуждаются в ненадежных терминологических и концептуальных интерпретациях.

Выводы всех без исключения разделов настоящего исследования показывают, что все найденные закономерности определяются самой природой языка – системы, существующей для того, чтобы хранить накопленные человеком и человечеством категоризации и передавать информацию в коммуникативных процессах.

В процессах порождения и восприятия речи носители языка постоянно передают и воспринимают новую информацию и постоянно оказываются в ситуациях необходимости принятия решения о релевантности этой информации и способах ее передачи. Все это определяет не жесткий, не полностью структурированный, динамический характер самой языковой системы. Сказанное касается не только самых высоких уровней языка, но и грамматики, в частности – морфологии. Грамматическое описание должно и может быть построено с учетом этих положений. Некоторые компоненты такого описания, как мне представляется, сформированы в настоящем исследовании» [Русакова 2009].

Литература

Антонов А. В., Ягунова Е. В. 2010. Охват содержимого информационных потоков путем анализа сверток текстов // *Материалы XII Всероссийской научной конференции RCDL'2010 «Электронные библиотеки : перспективные методы, технологии, электронные коллекции» (Казань, 13 –17 октября 2010 года)*. Казань, (в печати)

Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю. 2008. *Словарь глагольной сочетаемости непредметных имен русского языка* (электронный ресурс). Доступен для скачивания по адресу: http://dict.ruslang.ru/abstr_noun.php

Богданов С. И., Рыжова Ю. В. 1997. *Русская служебная лексика. Сводные таблицы*. СПб.

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

Касевич В. Б. 1988. *Семантика. Синтаксис. Морфология*. М.

Касевич В. Б., Ягунова Е. В. 2006. Корпуса письменных текстов и моделирование восприятия речи // *Вестник СПбГУ. Серия 2.* 2006, Вып.3.

Касевич В. Б., Ягунова Е. В. 2004. Перцептивный словарь взрослых и детей // *Проблемы социо- и психолингвистики. Сборник статей. Вып.6.* Пермь.

Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю. Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов. // *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной*

Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). *Вып. 9 (16)*. М.: Изд-во РГГУ. С. 181-185.

Копотев М. В. 2008. *Принципы синтаксической идиоматизации*. Хельсинки: Helsinki University Press.

Кустова Г. И. 2008. *Словарь русской идиоматики. Сочетания слов со значением высокой степени* (электронный документ). Доступен для скачивания по адресу: http://dict.ruslang.ru/magn.php

Мельчук И. А. 1960 О терминах «устойчивость» и «идиоматичность» // *Вопросы языкознания 4*. С. 73-80.

Пивоварова Л. М. 2010. Устойчивые конструкции, характеризующие тексты СМИ // *Прикладная и математическая лингвистика: Материалы секции XXXIX Международной филологической конференции*, СПб, (в печати).

Пивоварова Л.М., Ягунова Е. В. 2010 Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов (предварительные наблюдения) // *Материалы Симпозиума "Терминология и знание"* (Москва, май 2010 г.). М. (в печати)

Рогожникова Р. П. 2003. *Толковый словарь сочетаний, эквивалентных слову*. М.,

Русакова М. В. 2009 *Речевая реализация грамматических элементов русского языка*. Автореферат диссертации на соискание степени доктора филологических наук. СПб., 2009

Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А., 2008 *Статистический словарь языка русской газеты (1990-е годы)*. М.

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

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

Ягунова Е.В., Пивоварова Л.М. 2010б. Извлечение и классификация коллокаций на материале научных текстов. предварительные наблюдения // *V Международная научно-практическая конференция "Прикладная лингвистика в науке и образовании" памяти Р.Г. Пиотровского (1922-2009) : Материалы.* СПб. С. 356-364

Barlow M., Kemmer S. (eds) 2000. *Usage-based models of language.* Stanford, Calif.: CSLI Publications.

Boulis C. 2002. Clustering of Cepstrum Coefficients Using Pairwise Mutual Information. Tehnical Report EE516, University of Washington, Seattle.

Barðdal J. 2001 *Case in Icelandic – A Synchronic, Diachronic and Comparative Approach.* [Doctoral Dissertation]. Lundastudier i Nordisk språkvetenskap A 57. Department of Scandinavian Languages, Lund 2001.

Croft W. 2001. Radical *Construction Grammar. Syntactic theory in typological perspective.* Oxford: Oxford University Press.

Croft W., Cruse A. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Daudaravicius V. 2010. Automatic identification of lexical units. // *Computational Linguistics and Intelligent text processing CICling-2009*, Meksikas, Meksika.

Fillmore Ch. J., Kay, P.. 1993. *Construction Grammar Coursebook.* Manuscript, University of California at Berkeley Department of linguistics.

Fillmore Ch. J. 1999. Inversion and constructional inheritance.// Webelhuth G., Koenig J., Kathol A. (eds.). *Lexical and constructional aspects of linguistic explanation.* Stanford, Ca: CSLI. 113-128.

Fillmore Ch., J., Kay P., O'Connor M. C. 1988. Regularity and idiomaticity in grammatical constructions: The case of Let alone. // *Language 64, 3.* 501-538.

Firth, J.R.: 1957. *Papers in Linguistics 1934–1951*. London.

Firth, J.R.: 1968.*Selected Papers of J.R. Firth, 1952–1959.* London.

Fried M., Östman J. 2004. Construction Grammar: a thumbnail sketch.// Fried M., Östman J. (eds.). *Construction Grammar in a cross-language perspective.* 11-86. Amsterdam: John Benjamins.

Goldberg A. 1995. *Constructions. A Construction Grammar approach to argument structure.* Chicago: University of Chicago Press.

Goldberg A. 2006. *Constructions at Work: the nature of generalization in language*. Oxford University Press

Halliday M. 1966. Lexis as a Linguistic Level. // Bazell, C., Catford, J., Halliday, M., and Robins, R. (eds.). *In Memory of J. R. Firth.* Longman, London

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

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

Masini F. 2005. Multi-word Expressions between Syntax and the Lexicon: the case of Italian Verb-particle Constructions. SKY // *Journal of Linguistics 18.* 145-173

Mel'chuk I.A. 1995 Phrasemes in Language and Phraseology in Linguistics // *Idioms: Structural and Psychological perspectives*. Hillsdale, New Jersy 1995, 167-232

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

Su K.-Y., Wu M.-W., Chang J-S. 1994. A Corpus-based Approach to Automatic Compound Extraction. // *Proceedings, 32nd Annual Meeting of the ACL*. Las Cruces, NM, ACL, 242-247.

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

Tadić M., Šojat K. 2003. Finding multiword term candidates in Croatian. // *Proceedings of IESL2003 Workshop*. Borovets, Bulgaria. 102–107.