

УДК 811.161.1; 811.374; 81.322.2
DOI: 10.17223/22274200/21/2

М.В. Хохлова

АТРИБУТИВНЫЕ КОЛЛОКАЦИИ В ЗОЛОТОМ СТАНДАРТЕ СОЧЕТАЕМОСТИ РУССКОГО ЯЗЫКА И ИХ ПРЕДСТАВЛЕНИЕ В СЛОВАРЯХ И КОРПУСАХ ТЕКСТОВ¹

Рассматривается явление ограниченной сочетаемости и ее представление в словарях и корпусах русского языка. Приведен обзор лексикографических проектов и проанализировано, как описаны устойчивые словосочетания в толковых и специализированных словарях, насколько их данные совпадают между собой. Также обсуждаются следующие вопросы: в какой степени словарные коллокации отражены в корпусах текстов (НКРЯ и ruTenTen), и есть ли зависимость между количеством словарей, в которых зафиксировано словосочетание, и его частотностью.

Ключевые слова: *коллокации, устойчивая сочетаемость, атрибутивные словосочетания, русский язык, словари, корпуса текстов, база данных*

Введение

Изучению лексической сочетаемости при помощи статистических методов посвящено большое количество работ. В последнее время наблюдается бурное развитие разных подходов для извлечения коллокаций, главным образом включающих использование данных о совместной встречаемости и лингвистические фильтры. Результаты автоматического извлечения коллокаций представлены списками n-грамм (обычно биграмм), сопровождающимися числовыми коэффициентами, в которых могут быть репрезентированы единицы разной степени устойчивости: фразеологизмы («бить баклуши»), неоднословные единицы («без сомнения»), устойчивые словосочетания («принять решение»), именованные сущности («Красная площадь»). На русскоязычном материале подобные исследования проводились разными авторами (см., например, [1–4]). В последнее время стали использоваться методы машинного обучения применительно к задаче

¹ Исследование выполнено за счет гранта Российского научного фонда (проект № 19-78-00091).

автоматического выявления словосочетаний [5]. Задача создания списка высокочастотных коллокаций при помощи автоматических методов на основе большого корпуса текстов ставилась в исследовании [6], результаты работы применялись к снятию семантической омонимии.

В связи с развитием подобных методов возникают следующие вопросы: 1) каким образом можно проверить автоматически собранные данные и, следовательно, их оценить; 2) как можно оценить сами использованные методы. Соответственно, необходим некоторый золотой стандарт, который можно было бы применить как при проверке самих результатов, так при оценке работоспособности различных алгоритмов. Следовательно, существует потребность в ресурсах, которые содержали бы выверенные данные. При этом недостаточно использовать сведения только из одного источника (более того, вероятно, для русского языка пока не существует единого ресурса, в котором были бы собраны устойчивые словосочетания в большом объеме). Дополнительно упомянем некоторую сложность при использовании бумажных словарей [7], также рядом авторов отмечается несколько непоследовательный характер отбора и представления словарного материала [8, 9]. Проблема отбора словосочетаний является весьма важной в лексикографии, при этом не только для одноязычных словарей, и даже представляет собой «наиболее спорную и уязвимую часть практически каждого двуязычного словаря» [10. С. 61]. В работе [11] авторы указывают на сложность отбора примеров из корпуса, предлагая для данной задачи использовать списки коллокаций.

Обзор проектов

Информация о сочетаемости может быть представлена в традиционных (печатных, или «бумажных») словарях и в специализированных базах данных. Последние могут по своей сути также являться словарями с дополнительной информацией. Ниже мы остановимся на некоторых проектах, в которых с разной степенью проработанности дана информация о сочетаемости, при этом, безусловно, список ими не ограничивается. Необходимо сделать следующую оговорку: мы в большей степени хотели в него включить те современные проекты, которые или основываются на корпусном материале, или их данные получены при помощи автоматических методов. Тем не менее созда-

ние полноценного обзора подобных проектов (ресурсов) является отдельной задачей.

В основном работы, описывающие сочетаемость с использованием материала электронных корпусов текстов, на протяжении долгого времени затрагивали англоязычный материал. Комбинаторный словарь английского языка [12] был первой попыткой описать сочетаемость с помощью большого количества примеров. В нем описано более 75 тыс. словосочетаний и впервые разграничиваются лексические и грамматические коллокации. К первым относятся сочетания существительных, глаголов, прилагательных и наречий, в то время как вторая группа представлена словосочетаниями с предлогами. Словарь коллокаций «A Dictionary of English collocations» [13] – это одно из первых лексикографических произведений, при составлении которого использовались методы корпусной лингвистики. В нем перечислены все (по мнению автора) коллокации, встречающиеся в Брауновском корпусе. Также можно отметить словарь “The Pattern Dictionary of English Verbs”, который базируется на методологии Corpus Pattern Analysis, предложенной П. Хэнксом [14], и включает семантико-синтаксические шаблоны глагольного управления с иллюстрациями (словосочетаниями и предложениями). В настоящий момент речь также идет об автоматических словарях коллокаций [15].

Русскоязычная лексикографическая традиция имеет богатую историю, однако существует не так много проектов, посвященных сочетаемости в русском языке и основывающихся на корпусном материале или созданных с использованием автоматических методов. Уникальным проектом является Толково-комбинаторный словарь современного русского языка [16], в котором устойчивая сочетаемость описана при помощи лексических функций. Словарь устойчивых глагольно-именных словосочетаний русского языка [17] нацелен на изучающих русский язык и содержит 5 197 словосочетаний с 744 глаголами и 1 345 существительными, большинство из которых биграммы. Словарь коллокаций [18] был первым и единственным проектом, в название которого вынесено понятие, связанное с ограниченной сочетаемостью. В нем приводятся словосочетания для 512 заголовочных единиц, а также имеется англо-русский список ключевых слов. В Словаре сочетаемости слов русского языка [19] представлено 2 500 словарных статей для существительных, глаголов и прилагательных. Авторами делается различие между лексической и семантической сочетаемо-

стью, а также дается определение синтаксической сочетаемости как некоторой валентностной рамки. Существует уникальный лексикографический проект под руководством Ю. Д. Апресяна по созданию активного словаря русского языка [20], который включает обширную информацию о сочетаемости, отраженную отдельно в словарных статьях. Материал отлично структурирован и включает сведения о синтаксических актантах, коллокациях и конструкциях. Тем не менее, поскольку словари по-разному отражают сочетаемость и покрывают примеры, так важно рассмотрение отличных друг от друга источников.

Если говорить об электронных ресурсах, то для русского языка НКРЯ [21] предоставляет ряд инструментов (n-грамм поиск со статистической оценкой, списки устойчивых слов и словосочетаний, лексические графы), также на его основе были разработаны словари, в которых описана ограниченная сочетаемость. В словарь глагольной сочетаемости абстрактных существительных [22] вошли данные, полученные автоматически из синтаксически размеченного корпуса. В нем перечислена информация о более чем 10 000 словосочетаний следующих моделей: 1) N+V; 2) V+N; 3) V+ADJ+N. Авторами также используется аппарат лексических функций для выражения типовых смыслов. В словаре русской идиоматики [23] собраны сочетания со значением высокой интенсивности (наряду с другими он будет более подробно рассмотрен ниже).

Среди остальных ресурсов для русского языка можно назвать FrameBank [24], который включает описание валентностных рамок для глаголов и конструкций. База данных «Collocations, Colligations, Constructions» [25] предоставляет информацию о сочетаемости на основе НКРЯ и корпуса ruWac [26], также упомянем семантический словарь «Lexicograph» [27]. Отдельно следует назвать систему КроссЛексика [9], в которой содержится 1,75 млн словосочетаний. В словаре также отражены семантические связи: синонимия, антонимия, гипонимия и меронимия. В качестве лексикографической основы авторами был рассмотрен ряд двуязычных словарей, а также Словарь русского языка в четырех томах (Малый академический словарь) [28].

Далее будут более подробно рассмотрены некоторые из упомянутых словарных источников.

Постановка задачи

В настоящий момент нами ведется работа над проектом, целью которого является объединение и представление как традиционно описанных

устойчивых словосочетаний (в словарях, базах данных), так и тех, что были получены автоматически¹. Иными словами, речь идет о создании золотого стандарта сочетаемости, дополненного информацией из корпусов текстов. Словосочетания сопровождаются отсылками к словарям, в которых они были зафиксированы, а также ссылками на корпусные источники.

Как справедливо замечают некоторые авторы [23, 29], вопрос о разграничении словосочетаний разной степени связанности (если мы понимаем, что это свойство характерно в разной степени для всех словосочетаний) является до сих пор дискуссионным, поэтому «конкретные случаи идиоматической сочетаемости часто не получают однозначной квалификации, что отражается, в частности, в словарях» [23. С. 2]. Вопрос о том, какие словосочетания включать в компьютерный словарь, также обсуждается, например, в работе [30].

Сочетаемость лексической единицы вслед за [31] нами трактуется предельно широко: под ней понимается способность соединяться с другими единицами в речевой цепочке. Таким образом, нами будут рассматриваться словосочетания разной степени устойчивости. Например, по классификации [32]: идиомы («рабочая лошадь»), фразеологические словосочетания («телячий восторг»), речевые штампы («всего хорошего»), клише («минуту внимания»). Подобный подход к рассмотрению материала нам кажется оправданным, так как разные источники представляют данные разного рода. Соответственно, если бы мы ограничивались только одним типом, это привело бы к тому, что, во-первых, в базу данных попало бы лишь небольшое число примеров, во-вторых, ограниченное количество словосочетаний и их узконаправленность сделали бы подобный ресурс малоприменимым для последующего сравнения данных золотого стандарта с автоматически полученными.

Далее обсуждается та часть базы данных, в которую вошли материалы шести словарей (пяти печатных и одного электронного). В рамках нашего исследования поставлены две задачи. Во-первых, проанализировать, каким образом представлены устойчивые словосочетания в разных словарях, насколько они совпадают между собой. Во-вторых, исследовать, в какой степени коллокации определенного типа, которые отражены в словарях, могут быть найдены в корпусах текстов и, следовательно, проследить пересечение между «вручную» собранным материалом и современными корпусами.

¹ База данных коллокаций доступна онлайн: <https://collocations.spbu.ru>

Методика исследования

Первоначальный этап выполнения проекта был связан с отбором соответствующих лексикографических источников и с исследованием того, как в них представлена сочетаемость. В данной статье мы ограничимся атрибутивными двухсловными словосочетаниями, построенными по модели «прилагательное / причастие + существительное» (например, *заядлый охотник, каверзный вопрос, крепкая дружба* и др.), уделяя внимание их лексической структуре и их представлению в словарях и корпусах текстов.

Информация об ограниченной сочетаемости может быть найдена в толковых словарях, при этом не существует единой концепции описания данных, а также сами словосочетания могут быть рассмотрены с разной степенью подробности (просто перечислены в виде списка или при помощи полноценных словарных статей с соответствующим толкованием). Так, знак ромба указывает на устойчивые словосочетания в словаре и фразеологизмы в словаре МАС, в то время как в новом Большом академическом словаре русского языка (БАС) [33] для первых используется он же, а вторые вводятся при помощи знака тильды. По-разному представлены и сами словарные статьи: устойчивые словосочетания могут быть перечислены в зонах речений и цитат или выделяться иным образом. Существуют также иные словари, в которых представлены словосочетания, требующие отдельного описания. Некоторые из рассмотренных словарей были созданы для студентов-иностранцев и являются специализированными – в них более широко и подробно (чем в толковых словарях) показаны устойчивые словосочетания.

Таким образом, лексикографические источники, в которых может быть описана лексическая сочетаемость, представлены следующими типами: 1) толковые словари; 2) словари сочетаемости; 3) онлайн-словари и электронные базы данных¹. В нашей статье мы рассмотрим

¹ В случае русского языка и описания русскоязычного материала редко приходится говорить исключительно об онлайн-словарях (строго говоря, нам известны только словари Lingvo, а также некоторые другие иностранные проекты, ориентированные на языковые пары, включающие русский язык, например Оксфордские словари), так как сложно провести границу между ними и электронными базами данных. Таким образом, речь все же идет о списках словосочетаний, собранных в виде базы данных.

следующие словари, которые были использованы в качестве источников при создании золотого стандарта атрибутивных словосочетаний:

1. Большой толковый словарь русского языка: А–Я / сост., гл. ред. канд. филол. наук С.А. Кузнецов. СПб. : Норинт, 1998. (БТС) [34].
2. Словарь русского языка : в 4 т. / под ред. А.П. Евгеньевой. 4-е изд., стер. М. : Рус. яз., 1999. (МАС) [28].
3. Борисова Е.Г. Слово в тексте. Словарь коллокаций (устойчивых сочетаний) русского языка с англо-русским словарем ключевых слов. М. : Филология, 1995 [18].
4. Регина К.В., Тюрина Г.П., Широкова Л.И. Устойчивые словосочетания русского языка : учеб. пос. для студентов-иностранцев / под ред. Л.И. Широковой. М., 1980 [35].
5. Убин И.И. Словарь усилительных словосочетаний русского и английского языков. М. : Рус. яз., 1987 [36].
6. Кустова Г.И. Словарь русской идиоматики. Сочетания слов со значением высокой степени. М., 2008. URL: <http://dict.ruslang.ru/magn.php> [23].

Выше уже было отмечено, что словарные статьи отличаются как по своей структуре, так и по описанию лексической сочетаемости. Далее мы остановимся на принципах представления данных в лексикографических источниках, их особенностях и на их обработке, включающих формальные правила, которые были использованы для извлечения устойчивых выражений.

Анализ словарных данных

В России печатные словари широко распространены и являются, пожалуй, единственным типом словаря, а электронные словари (которые изначально разрабатываются для онлайн-доступа) занимают небольшую нишу¹. Среди рассмотренных нами источников единственным электронным ресурсом является словарь русской идиоматики [23], в то время как остальные доступны только в бумажном виде.

Извлечение данных из печатных словарей основывалось на структурных выделениях, принятых в словарных статьях. Отметим, что, как оказалось, в рассмотренных словарях (толковых и сочетаемости) зачастую содержится больше словосочетаний, чем то количество, ко-

¹ Согласно опросу лексикографов Elexis, только около 20% реализуемых ими проектов будут доступны в печатном виде [37].

торое указывается в предисловии. Например, в Словаре усилительных словосочетаний русского и английского языков [36] указано, что он содержит 10 400 единиц, при этом нами было извлечено на начальном этапе 12 894 словосочетания.

Толковые словари

В ходе исследования были рассмотрены два толковых словаря (МАС и БТС).

Малый академический словарь русского языка.

При анализе словарных статей в МАС на начальном этапе мы обратились к их заромбовой части, так как именно в ней представлена информация об ограниченной сочетаемости (о фразеологических оборотах и об устойчивых сочетаниях) и она оформлена как отдельный раздел. На материале МАС было получено 11 210 словосочетаний для 5 955 заголовочных слов (таким образом, при общем словнике словаря около 80 тыс. слов фразеологическую часть содержат только 7% словарных статей). Так, при лексеме «качественный» знаком ромба были отмечены два словосочетания: *качественный анализ* и *качественное прилагательное*. В самой словарной статье также содержатся примеры, иллюстрирующие три значения заголовочного слова: *качественные различия*, *качественный ремонт*, *качественные изменения*. Химический термин «качественный анализ» приводится как пример для второго значения слова «анализ» («определение состава и свойств какого-л. вещества, исследование их») в соответствующей статье, но не выделяется специальным образом. То же самое справедливо для словарной статьи «прилагательный»: *относительные прилагательные* и *качественные прилагательные* приводятся как иллюстрации. Граница между свободными или частотными (воспроизводимыми в речи) словосочетаниями, которые используются как примеры, и устойчивыми сочетаниями или терминами в заромбовой части также не всегда четко проводится авторами словаря. Например, для «абстрактный» в качестве типичного словосочетания указано *абстрактное понятие*, в то время как *абстрактное искусство* приводится за ромбом.

Большой толковый словарь русского языка.

БТС является преемником двух основополагающих проектов в отечественной лексикографии – Большого и Малого академических словарей – и основывается на том же материале. Нами также была

проанализирована зарембовая часть словарных статей, из которой было извлечено 8 205 словосочетаний. Авторы подчеркивают, что в данном разделе представлены фразеологизмы, в то время как «устойчивые словосочетания, сохраняющие ощутимую, ясную связь с каким-либо значением толкуемого слова, помещаются при данном значении в блоке иллюстрации» [34. С. 18]. Результаты показали, что не менее половины примеров совпадают с данными МАС. Например, *абсолютный ноль, бросить свет, придержать язык*. Можно также отметить, что ряд примеров являются устаревшими, книжными или сниженными по стилю. Например, *панургово стадо, воровские грамоты, отставной козы барабаничик*.

Словари сочетаемости

Как уже отмечалось выше, часть словарей, в которых представлена ограниченная сочетаемость, нацелена на иностранцев, изучающих русский язык. Этим объясняется специфика отбора материала и его отражение в источниках.

Словарь усилительных словосочетаний русского и английского языков.

В словаре [36] приводятся русские и английские усилительные словосочетания. Анализ, проведенный автором словаря, показал, что «усилительные словосочетания обладают высокой частотностью в обоих языках в самых разных текстах, особенно общественно-политических, где они встречаются в среднем через 80–100 слов» [36. С. 5]. Объем текстов составил 5 млн словоупотреблений для каждого языка. Словарь содержит прямую и реверсивную части, каждая из которых в свою очередь разделяется на русско- и англоязычную. Заглавным словом в прямом словаре является усиливаемое слово. Например, в словарной статье «экзамен»¹ (см. рис. 1) приводятся следующие коллокаты: *важный, ответственный, решающий, серьезный, сложный* и др. Предусмотрены пометы для обозначения бранной, разговорной, устаревшей и другой лексики.

¹ Далее нами будут по возможности рассмотрены словарные статьи для одних и тех же лексем из разных словарей, чтобы облегчить их сравнение. На рис. 1–3 сохранено представление словарных статей в печатных словарях, в том числе разбивка по строкам и выделение шрифтом.

ЭКЗАМЕН, -а, м – examination 1 важный, ответственный, решающий, серьёзный, сложный, строгий, суровый, трудный, тяжёлый
см. тж. экзаменовать, испытать, испытание, проверить, проверка

Рис. 1. Словарная статья для лексемы «экзамен»
 в Словаре усилительных словосочетаний русского и английского языков

В реверсивном словаре в качестве заглавных слов используются слова-усилители. В качестве синтагматических партнеров для прилагательного «твёрдый» перечислены следующие лексемы: *вера, власть, воля, гарантия, дисциплина, знание* и др.

Нами были рассмотрены словосочетания из реверсивной русско-язычной части словаря, так как она является более полной по сравнению с прямой частью. Заглавное слово или словосочетание (т.е. слово-усилитель) указывается при помощи прописных букв и полужирного шрифта, в то время как через запятую списком приводятся слова, для усиления которых может быть использовано данное заглавное слово. Словарные статьи, обозначенные знаком //, были исключены из рассмотрения, так как указывают на синонимы для заголовочных слов (в реверсивной части они даны в обратном порядке), объединяющие в своем значении значение самого слова и слова-усилителя. Например, «//ГУЛЛИВЕР высокий»; «//ОБОЖАТЬ любить»; «//ЭРУДИЦИЯ знания».

После проверки и удаления ошибок распознавания итоговый список составили 12 835 словосочетаний, из них 7 228 относятся к атрибутивным. Всего в атрибутивных словосочетаниях присутствуют 1 558 существительных и 717 прилагательных. Одному существительному соответствует в среднем 4,64 словосочетаний, в то время как на одно прилагательное приходится около 10 коллокаций. К наиболее частотным существительным относятся следующие: *успех* (41)¹, *тоска* (38), *борьба* (38), *сила* (36), *любовь* (34), *интерес* (32), *холод* (30), *мороз* (30), *ненависть* (29), *рост* (27), *радость* (27), *красота* (27), *ум* (24), *восторг* (24), *талант* (23), *победа* (23) и *значение* (23). При 742 существительных (47,63%) указано только одно прилагательное, которое с ними сочетается.

209 прилагательных имеют только один коллокат-существительное (т.е. 29,15% от всех прилагательных, зафиксированных в сло-

¹ Здесь и далее в скобках указывается число коллокаций с данной лексемой.

варных статьях). В качестве наиболее частотных прилагательных можно указать следующие: *большой* (340), *огромный* (251), *страшный* (216), *глубокий* (209), *полный* (166), *ужасный* (154), *сильный* (131), *колоссальный* (100), *поразительный* (86), *громадный* (83), *крупный* (79), *серьёзный* (77) и *широкий* (71).

Словарь коллокаций.

Список, составленный на материале словаря коллокаций [18] содержит 3 290 словосочетаний. Коллокации приводятся в нем при помощи заглавных букв, дополнительно отмечена семантическая информация (см. рис. 2).

ЭКЗАМЕН

1. ПРОВОДИТЬ/провести ЭКЗАМЕН какой/ по чему у кого.

Экзамен проводил молодой преподаватель. ПРИНИМАТЬ/принять ЭКЗАМЕН какой/ по чему у кого ПРОФЕССОР принимал экзамен по философии.

2. СДАВАТЬ/сдать ЭКЗАМЕН какой/ по чему кому Он сдал все экзамены на пять.

ДЕРЖАТЬ нет св пас ЭКЗАМЕН по чему

Завтра он держит самый важный экзамен.

3. ЭКЗАМЕН ИДЕТ нет св по чему/ какой

В институте или выпускные экзамены. ЭКЗАМЕН ПРОХОДИТ/ пройдет

Экзамены прошли незаметно. 8.1. ВЫДЕРЖИВАТЬ/выдержать нет пас

ЭКЗАМЕН какой/по чему Он выдержал экзамен и был принят.

9.1. ПРОВАЛИВАТЬСЯ/провалиться НА ЭКЗАМЕНЕ какой/по чему

Школьник провалился на экзамене по математике.

Рис. 2. Словарная статья для лексемы «экзамен» в Словаре коллокаций

Цифры отражают один из 11 стандартных смыслов, являющихся модификациями лексических функций, а также их сочетания. Например, цифра 1 означает «действие или состояние главного (активного) участника ситуации» [18. С. 5] («проводить экзамен», «провести экзамен», «принимать экзамен», «принять экзамен»), в то время как цифра 2 передает «действие или состояние второстепенных участников ситуации» [18. С. 5] («сдавать экзамен», «сдать экзамен», «держать экзамен»). Для извлечения данных при обработке словаря так же, как и в случае словаря [36], использовались шрифтовые выделения. Итоговый перечень при рассмотрении видовых пар как отдельных записей составил 5 334 единицы. В него вошли как коллокации, которые приведены в словарных статьях заглавными буквами, так и те единицы, которые были выделены нами вручную на основе анализа

иллюстративных примеров (в них коллокации не были отмечены специальным образом). Например, *вероломное нападение, выпускной экзамен, производительность труда, согревающий компресс*. Подобные словосочетания (всего 230 единиц) были дополнительно отмечены знаком * как содержащиеся в словаре, но при этом не указанные авторами как коллокации.

Всего в словаре было найдено 606 атрибутивных словосочетаний, в которых представлено 271 существительное. Таким образом, в среднем на одно существительное приходится 2,24 словосочетаний. К существительным, имеющим от 6 до 10 словосочетаний, относятся следующие: *положение* (10), *память* (9), *мысль* (9), *влияние* (9), *характер* (8), *интерес* (8), *обстановка* (7), *место* (7), *внимание* (7), *взгляд* (7), *отношение* (6), *образование* (6), *время* (6), *впечатление* (6), *вопрос* (6) и *борьба* (6). 307 прилагательных образуют атрибутивные коллокации, при этом они имеют более ограниченную сочетаемость по сравнению с существительными. Одно прилагательное в среднем зафиксировано в двух словосочетаниях. К наиболее частотным из них относятся следующие: *глубокий* (31), *высокий* (27), *острый* (16), *широкий* (15), *большой* (14), *тяжелый* (13), *полный* (12) и *низкий* (12).

Словарь устойчивых словосочетаний русского языка.

Словарь устойчивых словосочетаний русского языка [35] предназначен для иностранных учащихся и содержит около 3 000 устойчивых словосочетаний. При описании данного класса единиц авторы ориентируются на фразеологические сочетания в понимании В.В. Виноградова [38] и фразеологические выражения в трактовке Н.М. Шанского [39]. Выбор тем был обусловлен программой по русскому языку для студентов-иностранцев, материалом послужили учебники и учебные пособия для студентов-иностранцев, а также газеты и научно-популярные журналы. Анализ показал, что примеры содержат большое количество устаревшей лексики (например, *разрядка напряженности, трудящиеся массы*)¹, а также вхождение одних и тех же словосочетаний несколько раз в разных словарных статьях (например, при лексеме «видный» указано только одно словосочетание *видное место*, при этом дается отсылка к статье «место»).

¹ Сами авторы указывали на тот факт, что в словаре присутствует большое количество составных наименований, отражающих явления действительности (*общественная собственность, социалистическая собственность, идеологическая борьба* и др.) [35. С. 13].

Устойчивые атрибутивные словосочетания представлены в словаре гнездовым способом при заголовочном слове – прилагательном (в большинстве случаев) или существительном (см. рис. 3).

ЭКЗАМЕН ○ Держать экзамен. См. держать Идёт экзамен. См. идти Принимать экзамен. См. принимать. Проводить экзамен. См. проводить Сдавать экзамен. См. сдавать
--

Рис. 3. Словарная статья для лексемы «экзамен»
в Словаре устойчивых словосочетаний русского языка

Иллюстративные примеры нами не рассматривались, так как синтагматические партнеры лексем достаточно полно представлены в самих словарных статьях. Всего было извлечено 608 атрибутивных словосочетаний, построенных по модели «прилагательное/причастие + существительное»¹. В данном списке было зафиксировано 308 существительных. Одному существительному соответствует около двух словосочетаний. К наиболее частотным существительным относятся: *сила* (19), *связь* (16), *вопрос* (14), *роль* (12), *движение* (12), *интерес* (9), *жизнь* (9), *место* (8), *задача* (8), *проблема* (7), *время* (7) и *борьба* (7). В сводном списке словосочетаний зафиксировано 155 прилагательных, на каждое из которых приходится в среднем около четырех словосочетаний. К наиболее частотным прилагательным относятся следующие: *глубокий* (43), *высокий* (37), *широкий* (33), *большой* (24), *яркий* (15), *серьёзный* (13), *острый* (13), *общественный* (13), *крупный* (13), *экономический* (12), *тяжёлый* (12), *полный* (12), *низкий* (12), *народный* (12), *политический* (11), *горячий* (11) и *общий* (10).

Словарь русской идиоматики.

Словарь русской идиоматики [23] является наиболее современным из всех рассмотренных, при этом он доступен в электронном виде, что существенно облегчило его обработку. Словник основан на данных НКРЯ и бумажных словарей и включает широкий круг словосо-

¹ В словаре также содержатся атрибутивные словосочетания, построенные по моделям: 1) «существительное + существительное в родительном падеже» (*взрыв смеха, круг вопросов, дело мира*); 2) «существительное + предлог + существительное в косвенном падеже» (*право на самоопределение, борьба за мир, воля к борьбе*).

четаний: фразеологические единицы, коллокации, идиоматические выражения и семантически мотивированные свободные словосочетания. В словаре представлены сочетания слов со значением высокой степени и разных синтаксических моделей, например, *крупный авторитет*, *полностью убедить*, *абсолютно верно*, *безумно трудно*, *удивительно смешной* и др. В качестве степенных слов включены наречия и прилагательные. Из словаря было извлечено 10 150 записей, из них 6 935 составляют атрибутивные словосочетания.

Всего в списке представлено 1 888 существительных, на каждое из которых в среднем приходится около четырех коллокаций. Наиболее частотные существительные в основном передают чувства: *сила* (56), *успех* (46), *тоска* (42), *радость* (35), *любовь* (31), *страх* (29), *ненависть* (29), *красота* (27), *усилие* (26), *страсть* (26), *желание* (26), *восторг* (26), *ярость* (25), *холод* (25), *интерес* (25) и *боль* (25). Всего 427 прилагательных, индекс составляет 16,24 словосочетания на одно прилагательное. То есть в данном словаре наиболее полно представлена сочетаемость, на одно прилагательное приходится максимальное число словосочетаний по сравнению с другими рассмотренными словарями. К наиболее частотным прилагательным относятся следующие: *большой* (342), *великий* (235), *огромный* (230), *глубокий* (197), *настоящий* (182), *полный* (165), *сильный* (152), *крупный* (143), *высокий* (131), *значительный* (120), *абсолютный* (113), *страшный* (110), *колоссальный* (99), *необыкновенный* (98), *невероятный* (91), *поразительный* (85), *совершенный* (80), *исключительный* (79), *максимальный* (69) и *крайний* (60).

Объединение словарных данных

Следующий этап работы заключался в том, чтобы объединить коллокации из разных словарей и проанализировать пересекающиеся данные (насколько нам известно, это первая попытка сравнить таким образом словарный материал). На начальной стадии было извлечено более 20 тыс. коллокаций¹. Далее результаты были лемматизированы при помощи морфоанализатора MyStem [40]. Таблица 1 демонстрирует количество атрибутивных коллокаций, найденных в каждом источнике. Больше всего словосочетаний данного синтаксического типа

¹ Данные словосочетания мы будем называть *словарными коллокациями*.

было извлечено из словарей усилительных словосочетаний [36] и русской идиоматики [23].

Таблица 1

Количество извлеченных коллокаций по словарям

Борисова, 1995	Кустова, 2008	Убин, 1987	МАС	Регинина, Тюрина, Широкова, 1980	БТС
606	6935	7228	3278	608	1384

После фильтрации данных и удаления одинаковых записей начальный список был сведен к 15 101 единице.

Далее было произведено сравнение словарей между собой (см. табл. 2). Анализ показал, что словари МАС и БТС демонстрируют минимальное пересечение со словарями сочетаемости. Например, объединенный список МАС и словаря [23] содержит более 10 тыс. позиций, и только 55 из них представлены в обоих источниках (менее 1% всего списка). Это можно объяснить тем, что в словаре русской идиоматики описаны коллокации со значением высокой степени, в то время как МАС ориентирован на всестороннее представление лексики в целом и также делает акцент на фразеологических единицах.

Таблица 2

Попарное сравнение словарей

	Борисова, 1995	Кустова, 2008	Убин, 1987	МАС	Регинина, Тюрина, Широкова, 1980	БТС
Борисова, 1995		173	210	30	156	14
Кустова, 2008			3 465	55	154	11
Убин, 1987				45	199	15
МАС					29	745
Регинина, Тюрина, Широкова, 1980						3
БТС						

При попарном сравнении словарей следующие две пары показывают наибольшее совпадение: 1) словари усилительных словосочетаний [36] и русской идиоматики [23]; 2) МАС и БТС. Наибольшее пересечение коллокаций из первой пары словарей может быть объяснено значительным объемом извлеченных примеров по сравнению с другими словарями и, как следствие, наибольшим совпадением. В случае второй пары источников мы видим то, что уже наблюдали при предварительной обработке: согласно введению к БТС он был создан на материале МАС. Учитывая число выделенных коллокаций из БТС, объем пересечения оказывается около 54%.

Нами была введена метрика под названием «словарный индекс» для того, чтобы обозначить количество словарей, в которых представлено то или иное словосочетание. Большие значения данного индекса предполагают, что коллокация воспроизводима в речи довольно часто и, следовательно, должна быть выучена (если мы говорим об изучающих русский язык). Теоретически, максимум равен количеству рассмотренных словарей, т.е. в нашем случае 6, но на практике максимальное число словарей, в которых коллокация была зафиксирована, оказалось равным 4. В табл. 3 дан обзор результатов.

Таблица 3

Словарный индекс

Словарный индекс	Количество коллокаций
4	61
3	243
2	4 203
1	10 594

Тот факт, что ни одно из словосочетаний не представлено одновременно во всех словарях, может свидетельствовать о субъективности отбора примеров для словарей, а также об их недостаточном объеме¹. Максимальное пересечение данных (51 из 61 словосочетания с данным словарным индексом, т.е. 83,61%) наблюдается между словарями коллокаций [18], русской идиоматики [23], усилительных слово-

¹ Словари сочетаемости для других языков охватывают более обширный материал. Например, можно отметить Оксфордский словарь коллокаций для английского языка [41], который насчитывает 250 тыс. единиц, или словарь для немецкого языка [42], в котором приводится 192 тыс. словосочетаний.

сочетаний [36] и устойчивых словосочетаний [35]. К этим коллокациям относятся, например, сочетания с прилагательными *глубокий* (16) (*глубокая благодарность, глубокий интерес, глубокое удовлетворение*), *острый* (7) (*острая борьба, острая дискуссия, острая полемика*) и *широкий* (6) (*широкий выбор, широкая известность, широкое сотрудничество*).

Как и ожидалось, наиболее многочисленной оказалась группа коллокаций со словарным индексом 1. При этом данные словосочетания зафиксированы во всех шести словарях (см. табл. 4).

Таблица 4

Количество уникальных словосочетаний в словарях

Показатель	Борисова, 1995	Кустова, 2008	Убин, 1987	МАС	Регинина, Тюрина, Широкова, 1980	БТС
Количество словосочетаний	277	3316	3626	2445	296	634

Две трети от общего объема уникальных словосочетаний представлено в словаре русской идиоматики [23] и в словаре усилительных словосочетаний [36], что может быть объяснено их большим объемом, при этом МАС содержит наибольшее количество подобных коллокаций от общего объема словаря (74,59%).

Анализ коллокаций

Мы проанализировали ключевые слова и коллокаты (т.е. главные и зависимые единицы), которые присутствуют в объединенном списке.

Ключевые слова.

В целом коллокации содержат 3 606 существительных, из которых 1 550 (около 43%) встречаются только один раз, т.е. формируют только одну коллокацию. Оставшиеся существительные имеют от 2 до 97 коллокаций. Топ-50 включает те, которые отражают чувства и абстрактные понятия, например, *сила* (97), *успех* (59), *борьба* (55), *тоска* (54), *любовь* (49), *интерес* (46), *дело* (43), *болезнь* (42), *радость* (40), *память* (40), *красота* (38), *значение* (37), *чувство* (36), *система* (36), *ненависть* (36), *ум* (35), *страсть* (34), *роль* (34), *холод* (33), *усилие* (32). В словаре КроссЛексика иные существительные имеют

наибольшее количество определений [9]: *человек, лицо, работа, глаза, женщина, взгляд, вид, режим, голос*. Для сравнения отметим, что среди примеров в системе КроссЛексика есть существительные, которые относятся к наиболее частотным по словарю [43]. В этом их отличие от собранных нами данных: последние в основном входят во вторую–четвертую сотню частотного списка.

Коллокаты.

Всего в список вошло 2 841 прилагательное, из них 1 551 (т.е. около 55%) является уникальным. Оставшиеся прилагательные принимают участие в разнообразных коллокациях, превышающих несколько сотен. Большая часть прилагательных, которые чаще всего являются определениями, имеют значение большой степени, интенсивности или размера. И тут это можно объяснить тем, что они представляют собой примеры из словаря [23]. Например, *большой* (461), *огромный* (309), *великий* (275), *глубокий* (265), *страшный* (260), *сильный* (217), *полный* (215), *настоящий* (184), *высокий* (183), *ужасный* (173), *крупный* (170), *значительный* (130), *колоссальный* (120), *необыкновенный* (119), *абсолютный* (116), *широкий* (99), *поразительный* (98), *громадный* (94), *невероятный* (91), *тяжелый* (89). Наблюдается совпадение с высокочастотными прилагательными, представленными в системе КроссЛексика и в частотном словаре русского языка [43].

Необходимо отметить, что в отличие от существительных прилагательные показывают большую вариативность при построении коллокаций. В среднем на одно прилагательное приходится 5,3 коллокаций, в то время как на одно существительное – 4,2 коллокаций.

Сравнение с корпусами текстов

Объединение словарных коллокаций из разных источников предполагает не только единый лексикографический формат, но и релевантность данных. При описании материала лексикографу необходимо выбрать те примеры, которые основаны на их репрезентативности в корпусах, покрытии в словарях и также пригодности для пользователей языка и их целей. В ходе исследования мы хотим проверить следующую гипотезу: большие частоты коллокаций в корпусе соответствуют высоким показателям вышеупомянутого словарного индекса. То есть мы ожидаем увидеть прямо пропорциональную зависимость между словарными и корпусными данными и, следовательно, положительную корреляцию между словарями и корпусами.

Для оценки выделенных словосочетаний относительно корпусов текстов были случайно выбраны 20 коллокаций из групп со словарными индексами 2, 3 и 4 (см. табл. 5).

Таблица 5

Примеры коллокаций из групп с разными словарными индексами

№	Словарная коллокация	Борисова, 1995	Кустова, 2008	Убин, 1987	МАС	Регинина, Тюрина, Широкова, 1980	БТС	Словарный индекс
1	адская боль	0 ¹	1	1	0	0	0	2
2	великое переселение	0	1	0	1	0	1	3
3	высокий результат	1	1	1	0	1	0	4
4	высокое мастерство	0	1	1	0	1	0	3
5	глубокая древность	0	1	1	0	1	0	3
6	горячая любовь	1	1	1	0	1	0	4
7	жгучий стыд	1	1	1	0	0	0	3
8	жизненный путь	1	0	0	1	1	0	3
9	звериная скука	0	1	1	0	0	0	2
10	звериная жестокость	0	1	1	0	0	0	2
11	львиная часть	0	0	0	1	0	1	2
12	острая нужда	1	1	1	0	1	0	4
13	острая дискуссия	1	1	1	0	1	0	4
14	путеводная нить	0	0	0	1	0	1	2
15	широкая известность	1	1	1	0	1	0	4

В качестве материала были использованы следующие корпуса: подкорпус со снятой омонимией объемом 6 млн слов и основной корпус объемом 321 млн слов НКРЯ [21], а также интернет-корпус ruTenTen объемом 14,5 млрд слов [44]. В таблицах 6–8 представлены относительные частоты отобранных коллокаций.

На основе вышеприведенных данных (см. табл. 6) мы видим, что даже для высокого словарного индекса результаты оказываются неоднородными. В подкорпусе со снятой омонимией не были найдены 4 словосочетания, хотя они присутствуют в нескольких словарях.

¹ При помощи цифр 1 и 0 обозначено наличие или отсутствие словосочетания в словаре.

Таблица 6

Коллокации со словарным индексом 4

№	словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	большая разница	1,33	2,06	1,85
2	большое значение	7,33	9,27	12,65
3	большой успех	5,16	7,61	4,92
4	высокий результат	1,00	0,53	3,83
5	высокий урожай	0,00	0,92	0,75
6	глубокое уважение	0,50	1,88	1,10
7	глубокое удовлетворение	0,83	0,55	0,34
8	глубокий смысл	2,17	1,52	1,23
9	горячая любовь	0,17	1,20	0,24
10	крепкая дружба	0,17	0,22	0,29
11	острая нужда	0,00	0,42	0,16
12	острая дискуссия	0,00	0,28	0,33
13	острая критика	0,33	0,25	0,21
14	полная свобода	2,17	4,76	2,24
15	слабое место	1,50	2,20	3,14
16	широкая известность	1,00	1,00	1,26
17	широкая дискуссия	0,17	0,15	0,15
18	широкая поддержка	0,00	0,25	0,39
19	широкий размах	0,50	0,72	0,31
20	яркий пример	2,50	2,83	4,84

Коэффициент ранговой корреляции Спирмена варьируется от 0,81 до 0,82, что свидетельствует о сильной положительной связи частот словосочетаний в каждой из трех пар корпусов. Средняя частота для коллокации в подкорпусе НКРЯ составила 1,34, в основном корпусе НКРЯ – 1,93, в корпусе ruTenTen – 2,01, различия между данными не являются статистически значимыми ($p > 0,05$ согласно критерию Фридмана). При этом словосочетания с лексемой «большой» (*большое значение* и *большой успех*) имеют существенно более высокие частоты.

Частоты словосочетаний со словарным индексом 3 (см. табл. 7) также показывают высокую связанность (коэффициент ранговой корреляции Спирмена принимает значения от 0,63 до 0,79), при этом отличия между ними на материале трех корпусов статистически значимы ($p < 0,05$ согласно критерию Фридмана). Согласно *post hoc* критерию Немея различия наблюдаются между данными основного корпуса НКРЯ и корпуса ruTenTen ($p < 0,05$). Таким образом, в случае словосочетаний с индексом 3 частоты более разнообразны.

Таблица 7

Коллокации со словарным индексом 3

№	Словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	большая беда	0,83	1,76	0,60
2	большая польза	0,50	2,42	1,13
3	большая важность	0,33	0,93	0,25
4	большая помощь	0,66	0,98	1,23
5	великое переселение	0,33	0,45	0,34
6	высокое мастерство	0,50	0,38	0,68
7	высокая требовательность	0,17	0,14	0,13
8	гигантский шаг	1,00	0,62	0,14
9	глубокая древность	1,83	1,80	1,62
10	глубокое влияние	0,33	0,19	0,17
11	жгучий стыд	0,50	0,23	0,04
12	жизненный путь	2,83	4,03	3,87
13	коренной интерес	0,50	0,20	0,14
14	лютая ненависть	0,83	0,45	0,21
15	лютый мороз	1,50	0,73	0,42
16	набитый дурак	0,00	0,13	0,01
17	последние известия	2,00	2,23	0,24
18	равные права	1,00	1,45	1,63
19	тесная дружба	0,50	0,81	0,16
20	тяжёлая задача	0,00	0,21	0,13

Для коллокаций, отобранных из двух словарей (табл. 8), мы видим, что 12 словосочетаний из 20 (60 %) не были найдены в шестимиллионном подкорпусе НКРЯ, а 3 из них не зафиксированы также в основном корпусе. Коэффициент корреляции Спирмена уменьшился для двух пар корпусов (0,63 для корпусов НКРЯ и 0,53 для корпуса со снятой неоднозначностью НКРЯ и ruTenTen) и увеличился для основного корпуса НКРЯ и ruTenTen (и стал равен 0,92). Таким образом, мы можем отметить почти идентичное ранжирование в последнем случае. Как и со словарным индексом 4, различия в частотах между корпусами для данных словосочетаний не являются статистически значимыми ($p > 0,05$ согласно критерию Фридмана).

Частоты коллокаций с разными словарными индексами (за исключением словарных индексов 3 и 4) отличаются в одних и тех корпусах (различия статистически значимы, $p < 0,05$ согласно критерию Краскела–Уоллиса). Таким образом, можно отметить, что частота встречаемости словосочетаний в корпусах неуклонно падает с уменьшением количества словарей, в которых эти словосочетания зафиксированы.

Таблица 8

Коллокации со словарным индексом 2

№	словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	адская боль	0,50	0,17	0,15
2	безграничная тоска	0,00	0,01	0,01
3	безмерная глубина	0,17	0,03	0,01
4	жгучая злость	0,00	0,00	0,01
5	жестокое наказание	0,00	0,71	0,21
6	звериная скука	0,00	0,00	0,01
7	звериная жестокость	0,00	0,07	0,04
8	исключительное многообразие	0,00	0,01	0,01
9	львиная часть	0,00	0,12	0,13
10	местный падеж	0,00	0,02	0,01
11	нервная система	9,66	9,46	17,72
12	неукротимая злоба	0,17	0,05	0,01
13	огромный диапазон	0,00	0,07	0,09
14	отчаянная храбрость	0,00	0,23	0,04
15	полный восторг	0,33	0,98	0,88
16	поразительная предосторожность	0,00	0,00	0,01
17	путеводная нить	0,17	0,40	0,17
18	тотальная слежка	0,00	0,04	0,08
19	ужасная гроза	0,17	0,11	0,02
20	цепная реакция	1,50	1,97	1,32

Далее были отобраны случайным образом 20 словосочетаний, которые присутствуют только в одном из словарей (т.е. их словарный индекс был равен 1), и была проверена их частотность по трем корпусам. Таблицы 9–14 демонстрируют полученные результаты.

Несмотря на то что словарь русской идиоматики [23] был составлен на основе НКРЯ, только половина коллокаций была зафиксирована в подкорпусе со снятой морфологической неоднозначностью (табл. 9). Словосочетания, полученные на основе данного словаря, характеризуются крайне низкими частотами в обоих корпусах и показывают минимальные значения по сравнению с другими словарями. В корпусах средние значения частот составляют 0,09, 016 и 0,10 ipm¹ для корпусов НКРЯ и ruTenTen соответственно (различия не являются значимыми, $p > 0,05$ согласно критерию Фридмана), также низкими

¹ ipm – количество употреблений на 1 млн слов корпуса (англ. instances per million words).

оказываются значения стандартного отклонения, т.е. можно предположить некоторую однородность атрибутивных словосочетаний в данном словаре.

Таблица 9

Коллокации из Словаря русской идиоматики

№	словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	всесторонний анализ	0,17	0,21	0,46
2	крайнее беспокойство	0,00	0,09	0,02
3	сказочное богатство	0,17	0,19	0,08
4	огромная величина	0,67	0,47	0,08
5	сплошной восторг	0,00	0,04	0,02
6	значительная высота	0,17	0,63	0,23
7	твердая дисциплина	0,00	0,06	0,02
8	большое изобилие	0,00	0,15	0,06
9	великая красота	0,00	0,12	0,03
10	невероятная ловкость	0,17	0,01	0,02
11	прирожденный педагог	0,00	0,05	0,01
12	истовая преданность	0,00	0,00	0,01
13	грандиозный проект	0,50	0,40	0,50
14	недюжинная сила	0,33	0,21	0,13
15	отчаянный смельчак	0,00	0,08	0,04
16	великая страсть	0,17	0,15	0,04
17	крупный торговец	0,17	0,18	0,09
18	необычайный успех	0,17	0,16	0,04
19	поразительная целеустремленность	0,00	0,00	0,01
20	прожженный шалопай	0,00	0,00	0,01

Словосочетания, извлеченные из словаря [36], также характеризуются низкими частотами в корпусах (табл. 10) и отсутствием статистически значимых различий ($p > 0,05$ согласно критерию Фридмана).

Словарь устойчивых словосочетаний [35] показывает наиболее высокие результаты для частот (табл. 11), т.е. можно предположить, что в данном источнике отражены более частотные коллокации по сравнению с другими словарями, например, *высшее образование* и *духовная жизнь*. Различия в частотах между корпусами статистически незначимы ($p > 0,05$ согласно критерию Фридмана).

Анализ данных словаря [18] (табл. 12) позволяет предположить, что извлеченные из него коллокации по своим частотным характеристикам занимают промежуточное положение: 8 словосочетаний не были зафик-

сированы в подкорпусе со снятой омонимией, и 5 коллокаций имеют только одно вхождение. Различия в частотах между корпусами статистически значимы ($p < 0,05$ согласно критерию Фридмана).

Таблица 10

**Коллокации из Словаря усилительных словосочетаний
русского и английского языков**

№	Словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	активная кампания	0,00	0,04	0,08
2	близкий приятель	0,00	0,43	0,06
3	глубокая критика	0,33	0,05	0,01
4	доблестный защитник	0,00	0,12	0,05
5	долголетний опыт	0,00	0,12	0,05
6	жестокий обстрел	0,00	0,07	0,01
7	колючий мороз	0,00	0,01	0,01
8	молниеносный удар	0,00	0,12	0,07
9	надёжный помощник	0,50	0,16	0,53
10	огромная очередь	0,83	0,43	0,70
11	основное условие	0,83	0,90	1,87
12	протяжный стон	0,50	0,27	0,05
13	решительный характер	0,17	0,23	0,06
14	стойкое сопротивление	0,00	0,05	0,03
15	существенный признак	0,00	0,57	0,57
16	томительное ожидание	0,50	0,83	0,27
17	утонченный вкус	0,17	0,16	0,21
18	чрезмерная гордость	0,00	0,02	0,02
19	штормовой ветер	0,33	0,18	0,32
20	язвительная насмешка	0,00	0,12	0,03

Таблица 11

Коллокации из Словаря устойчивых словосочетаний русского языка

№	Словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	внутренняя проблема	0,33	0,47	0,96
2	военная сила	1,83	2,82	1,51
3	высшее образование	13,33	7,85	19,34
4	генеральный план	1,17	1,00	3,79
5	глубокий лиризм	0,00	0,03	0,02
6	деловой человек	2,67	2,66	2,98
7	дружественное отношение	1,17	0,95	0,63
8	духовная жизнь	4,66	8,03	4,17
9	историческая дата	0,00	0,15	0,22

№	Словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
10	крупный капиталист	0,00	0,22	0,06
11	культурный уровень	0,50	1,32	0,91
12	материалистическая философия	0,17	0,20	0,08
13	международная обстановка	0,83	0,91	0,39
14	мирный договор	1,67	2,38	1,79
15	народный поэт	0,00	0,47	0,25
16	теплый прием	0,33	0,25	0,79
17	низкая стоимость	0,33	0,39	3,36
18	ответная речь	0,33	0,36	0,09
19	широкое признание	0,17	0,33	0,50
20	яркая личность	0,67	0,44	0,58

Таблица 12

Коллокации из Словаря коллокаций

№	Словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	всестороннее образование	0,00	0,07	0,04
2	въездная виза	0,33	0,21	0,39
3	грустная мысль	0,50	0,83	0,36
4	душевное равновесие	0,83	1,29	1,09
5	жаркий спор	0,17	0,57	0,41
6	заметный успех	0,17	0,19	0,25
7	замечательная возможность	0,00	0,07	0,71
8	клеветническая кампания	0,00	0,07	0,05
9	многолетняя дружба	0,00	0,29	0,12
10	нежная привязанность	0,00	0,19	0,06
11	необъяснимая грусть	0,00	0,02	0,01
12	положительная оценка	0,83	0,47	1,35
13	предупредительный сигнал	0,17	0,11	0,16
14	серьезная авария	0,00	0,26	0,52
15	сырая погода	0,50	0,50	0,37
16	твёрдая вера	0,50	0,66	0,21
17	удобный случай	4,66	4,92	1,53
18	чрезвычайное положение	0,17	0,87	1,42
19	чудесная идея	0,00	0,02	0,02
20	широкий рынок	0,17	0,04	0,20

Результаты подтверждают уже отмеченную выше схожесть данных в рассмотренных толковых словарях, хотя источники несколько отличаются по тому, как в них представлены уникальные словосочетания. Для коллокаций из БТС (табл. 13) распределение частот в кор-

пусах характеризуется выбросами и большим размахом значений (например, *акционерное общество, органическое вещество, почтовый ящик*).

Различия между корпусами в наблюдаемых частотах для коллокаций из толковых словарей так же, как и в большинстве примеров, не являются статистически значимыми ($p > 0,05$ согласно критерию Фридмана).

Таблица 13

Коллокации из БТС

№	Словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	акционерное общество	6,66	5,47	14,37
2	богемское стекло	0,17	0,04	0,06
3	бриллиантовая свадьба	0,17	0,01	0,07
4	волнистый попугай	0,00	0,01	0,24
5	гашёная марка	0,00	0,01	0,01
6	житейское дело	0,67	0,67	0,15
7	золотая молодёжь	0,33	0,67	0,36
8	красная книга	0,50	0,89	2,70
9	нашатырный спирт	1,00	1,12	0,73
10	органическое вещество	3,00	3,33	2,69
11	парниковый эффект	0,17	0,62	0,70
12	подземное царство	0,33	0,54	0,43
13	почётный гражданин	1,00	1,73	2,70
14	почтовый ящик	5,66	3,75	6,25
15	саблезубый тигр	0,17	0,23	0,16
16	сборная таблица	0,00	0,00	0,01
17	согласный звук	0,33	0,19	0,11
18	спускаемый аппарат	0,00	1,06	0,45
19	старая вешалка	0,00	0,01	0,01
20	экстраординарный профессор	0,00	0,29	0,08

Словосочетания из МАС характеризуются меньшими отклонениями от средних значений (табл. 14).

Обсуждение результатов

Словарные коллокации являются в статистическом смысле редкими (и даже уникальными в определенной степени) явлениями. Около 34% рассмотренных примеров отсутствуют в подкорпусе НКРЯ со

снятой омонимией, т.е. можно предположить, что объема в 6 млн словоупотреблений недостаточно для изучения явления ограниченной сочетаемости¹.

Таблица 14

Коллокации из МАС

№	Словарная коллокация	НКРЯ, снят.	НКРЯ, осн.	ruTenTen
1	бутылочный цвет	0,17	0,08	0,01
2	вегетативное размножение	0,00	0,08	0,23
3	военный поселенец	0,00	0,04	0,02
4	вольная земля	0,00	0,24	0,02
5	грибной дождь	0,17	0,09	0,03
6	действующая армия	0,83	3,88	0,96
7	жесткая вода	0,17	0,16	0,45
8	затяжной прыжок	0,00	0,12	0,08
9	зимний сад	0,50	1,20	3,00
10	конституционная монархия	0,33	0,60	0,44
11	кровеное давление	1,67	0,90	2,23
12	местное время	0,67	0,63	2,74
13	мягкий вагон	0,83	0,30	0,03
14	обратный билет	1,17	0,53	0,43
15	плоская стопа	0,17	0,01	0,03
16	поджелудочная железа	0,83	0,93	0,01
17	прогулочный шаг	0,17	0,11	0,13
18	растительный орнамент	0,17	0,17	0,44
19	частотный словарь	0,00	0,13	0,06
20	чистое небо	1,33	1,41	0,68

При этом около 12% проанализированных коллокаций отличаются также очень низкими частотами (менее 0,01 ipm) даже в сверхбольшом интернет-корпусе ruTenTen. Присутствие коллокаций в нескольких словарях указывает на их более высокие частоты и, следовательно, воспроизводимость в речи. Основной корпус НКРЯ содержит большее количество примеров, но они также имеют низкие частоты, а наблюдаемые различия между корпусами статистически незначимы для всех рассмотренных словарей ($p > 0,05$ согласно критерию Уилкоксона). Таким образом

¹ Строго говоря, частота, отличная от 0 и равная, например, 1 ipm (т.е. 6 примерам для подкорпуса со снятой неоднозначностью НКРЯ), является недостаточной, т.к. для описания лексической единицы согласно [45] необходимо в среднем 50 примеров для английского языка.

мы можем заключить, что увеличение объема корпуса важно и необходимо при изучении устойчивых словосочетаний.

Сравнение частотных данных для словосочетаний со словарным индексом 1 (т.е. отмеченных только в одном из словарей) показало, что в целом между ними нет статистически значимых различий. Тем не менее в ряде случаев они присутствуют – с увеличением объема корпуса также увеличивается количество словарей, словосочетания из которых отличаются. Для корпуса со снятой неоднозначностью НКРЯ различия между частотами словосочетаний были обнаружены в словаре русской идиоматики [23] и словаре устойчивых словосочетаний [35] (в этом случае $p < 0,05$ согласно критерию Краскела–Уоллиса и *post hoc* тесту). Основной корпус НКРЯ наряду с предыдущим примером продемонстрировал значимые различия для еще одной пары словарей: устойчивых словосочетаний [35] и усилительных словосочетаний [36]. Для корпуса *ruTenTen* наблюдается больше различий в частотах словосочетаний, представленных в разных словарях: в дополнение к упомянутым двум парам также наблюдаются значимые отличия для словарей русской идиоматики [23] и коллокаций [18].

За исключением нескольких коллокаций (*большое значение, большой успех, высшее образование и нервная система*), примеры оказываются низкочастотными на материале трех корпусов. Подтверждается гипотеза о том, что с уменьшением словарного индекса относительные частоты встречаемости коллокаций в корпусе убывают (исключение составляют уникальные коллокации в словаре устойчивых словосочетаний [35], частоты которых, наоборот, превышают остальные¹). Анализ результатов позволяет сделать вывод о том, что в словарь [35] вошли более частотные словосочетания по сравнению с другими источниками, в то время как словарь русской идиоматики [23] содержит наименее воспроизводимые единицы.

Оба рассмотренных толковых словаря показывают крайне небольшое пересечение с другими лексикографическими источниками. Это можно объяснить тем фактом, что словари направлены на описание разной лексики: например, в словаре русской идиоматики [23] присутствуют только словосочетания с высокой степенью, в то время как МАС и БТС нацелены на более полное представление лексики,

¹ Это может быть объяснено разнообразием типов словосочетаний в данных словарях и, следовательно, большим разбросом частот.

при этом в исследованной нами заромбовой части представлены фразеологизмы.

Заключение

Нами были рассмотрены атрибутивные словосочетания, характеризующиеся определенной устойчивостью, и их представление в ряде словарей русского языка. К сожалению, несмотря на обработку нескольких источников, объем выделенных данных пока является недостаточным, поэтому важно анализировать иные словари и лексикографические источники. Пересечение данных, отраженных в разных источниках, является крайне низким (в особенности между толковыми словарями и словарями сочетаемости). Всего было обработано более 15 тыс. примеров, из которых менее 0,5% содержится в четырех из шести рассмотренных словарей. Сочетания, полученные из разных источников, демонстрируют разную природу. Коллокации, которые отмечены в нескольких словарях, оказываются более частотными и, как следствие, могут быть предсказаны автоматическими методами. Однако в целом проанализированные примеры характеризуются низкой встречаемостью в корпусах (подавляющее число словосочетаний, извлеченных из словарей, являются уникальными, т.е. представлены только в одном источнике). Результаты показывают, что весьма важным является объем данных, а само явление устойчивой сочетаемости необходимо исследовать на материале больших корпусов.

Литература

1. 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.
2. Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2010. № 6. С. 30–40.
3. Pivovarova L., Kormacheva D., Kopotev M. Evaluation of collocation extraction methods for the Russian language // *Quantitative Approaches to the Russian Language* / ed. by M. Kopotev, O. Lyashevskaya, A. Mustajoki. London ; New York : Routledge, 2018. P. 137–157.
4. Khokhlova M. Similarity between the Association Measures: a Case Study of Noun Phrases // *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic*

Natural Languages Processing, RASLAN 2018 / ed. by A. Horák, P. Rychlý, A. Rambousek. Brno : Tribun EU, 2018. P. 21–27.

5. *Enikeeva E., Mitrofanova O.* Russian Collocation Extraction Based on Word Embeddings // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая – 3 июня 2017 г.). Вып. 16 (23) : в 2 т. М. : Изд-во РГТУ, 2017. URL: <http://www.dialog-21.ru/media/3908/enikeevaevmitrofanovaoa.pdf> (дата обращения: 09.07.2021).

6. *Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю.* Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка // Интернет-математика 2005. Автоматическая обработка веб-данных. М., 2005. С. 38–57.

7. *Василисина А.А., Зарибян М.С., Казакова П.Н., Сударикова Е.А.* Полисемия глаголов перемещения, движения и локализации (по результатам семантической разметки базы данных Активного словаря русского языка). URL: <http://www.dialog-21.ru/media/3471/vasilisina.pdf> (дата обращения: 09.07.2021).

8. *Азарова И.В., Синопальникова А.А., Яворская М.В.* Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог 2004» (Верхневолжский, 2–7 июня 2004 г.). М., 2004. С. 542–547.

9. *Большаков И.А.* КроссЛексика – большой электронный словарь сочетаний и смысловых связей русских слов // Компьютерная лингвистика и интеллектуальные технологии : труды междунар. конф. «Диалог 2009». Вып. 8 (15). М. : РГТУ, 2009. С. 45–50.

10. *Берков В.П.* Двухязычная лексикография. 2-е изд. М. : АСТ, 2004. 236 с.

11. *Atkins B.T.S., Rundell M.* The Oxford Guide to Practical Lexicography. Oxford U.P, 2008. 554 p.

12. *Benson M., Benson E., Ilson R.* The BBI Combinatory Dictionary of English: A Guide to Word Combinations. Amsterdam, Philadelphia : John Benjamins, 1986. 462 p.

13. *Kjellmer G.* A dictionary of English collocations: based on the Brown corpus: in three volumes. Oxford ; New York : Clarendon Press: Oxford University Press, 1994. 2304 p.

14. *Hanks P.* Mapping meaning onto use: a Pattern Dictionary of English Verbs // Proceedings AACL 2008, Utah. 2008. URL: <https://nlp.fi.muni.cz/projects/cpa/Pattern%20Dict%20Utah.ppt> (дата обращения: 09.07.2021).

15. *Pezik P.* Graph-Based Analysis of Collocational Profiles // Phraseologie im Wörterbuch und Korpus (Phraseology in Dictionaries and Corpora) / ed. by Vida Jesenšek and Peter Grzybek. ZORA 97. Maribor ; Bielsko-Biała ; Budapest ; Kansas ; Praha : Filozofská fakulteta, 2014. P. 227–43.

16. *Мельчук И.А., Жолковский А.К. и др.* Толково-комбинаторный словарь современного русского языка. Опыт семантико-синтаксического описания русской лексики. Вена : Wiener Slavistischer Almanach, 1984. 992 с.

17. *Дерибас В.М.* Устойчивые глагольно-именные словосочетания русского языка. М. : Русский язык, 1983. 256 с.

18. Борисова Е.Г. Слово в тексте. Словарь коллокаций (устойчивых сочетаний) русского языка с англо-русским словарем ключевых слов. М. : Филология, 1995. 150 с.

19. Денисов П.Н., Морковкин В.В. Словарь сочетаемости слов русского языка. 3-е изд., испр. М. : АСТ, 2002. 688 с.

20. Апресян В.Ю., Апресян Ю.Д., Бабаева Е.Э., Богуславская О.Ю., Галактионова И.В., Гловинская М.Я., Иомдин Б.Л., Крылова Т.В., Левонтина И.Б., Птенцова А.В., Санников А.В., Урысон Е.В. Проспект активного словаря русского языка. URL: http://san.ruslang.ru/prospect_theory.pdf (дата обращения: 09.07.2021).

21. Национальный корпус русского языка. URL: <http://ruscorpora.ru> (дата обращения: 09.07.2021).

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

23. Кустова Г.И. Словарь русской идиоматики. Сочетания слов со значением высокой степени. М., 2008. URL: <http://dict.ruslang.ru/magn.php> (дата обращения: 09.07.2021).

24. Lyashevskaya O. Bank of Russian Constructions and Valencies // Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, 2010. P. 1802–1805.

25. Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangerber R. CoCoCo: Online Extraction of Russian Multiword Expressions // The 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria). Sofia : INCOMA Ltd, 2015. P. 43–45.

26. Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Беласово, 25–29 мая 2011 г.). Вып. 10 (17). М. : РГТУ, 2011. С. 657–670.

27. База данных «Lexicograph». URL: <http://lexicograph.ruslang.ru> (дата обращения: 09.07.2021).

28. Словарь русского языка : в 4 т. / под ред. А.П. Евгеньевой. 4-е изд., стер. М. : Рус. яз., 1999.

29. Calzolari N., Fillmore Ch., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. Towards Best Practice for Multiword Expressions in Computational Lexicons // Proceedings of LREC – 2002. P. 1934–1940.

30. Лукашевич Н.В., Добров Б.В., Чуйко Д.С. Отбор словосочетаний для словаря системы автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии : труды международной конференции «Диалог–2008». М., 2008. С. 339–344.

31. Тестелец Я.Г. Введение в общий синтаксис. М. : РГТУ, 2001. 798 с.

32. Телия В.Н. Русская фразеология. Семантический, прагматический и лингвокультурологический аспекты. М. : Языки славянской культуры, 1996. 289 с.

33. Большой академический словарь русского языка % в 20 т. М. : Российская академия наук. Институт лингвистических исследований. Наука, 2004–.

34. *Большой толковый словарь русского языка: А–Я* / сост., гл. ред. С.А. Кузнецов. СПб. : Норинт, 1998. 1534 с.

35. *Регинина К.В., Тюрина Г.П., Широкова Л.И. Устойчивые словосочетания русского языка : учеб. пособие для студентов-иностранцев / под ред. Л.И. Широковой. М. : Рус. яз., 1980. 296 с.*

36. *Убин И.И. Словарь усиленных словосочетаний русского и английского языков. М. : Рус. яз., 1987. 306 с.*

37. *Kallas J., Koeva S., Langemets M., Tiberius C., Kosem I. Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs // Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal. Brno : Lexical Computing CZ, s.r.o., 2019. P. 519–536.*

38. *Виноградов В.В. Русский язык. М. : Высшая школа, 1972. 613 с.*

39. *Шанский Н.М. Фразеология современного русского языка : учеб. пособие для студ. филол. фак-тов. 3-е изд., испр. и доп. М. : Высшая школа, 1985. 160 с.*

40. *MyStem*. URL: <https://yandex.ru/dev/mystem/> (дата обращения: 09.07.2021).

41. *Oxford Collocations Dictionary for Students of English, 2nd edition / ed. by Colin MacIntosh. Oxford : Oxford University Press, 2009. 992 p.*

42. *Quasthoff U. Wörterbuch der Kollokationen im Deutschen. Berlin ; New York : Walter de Gruyter, 2011. 551 S.*

43. *Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М., 2009. URL: <http://dict.ruslang.ru/freq.php> (дата обращения: 09.07.2021).*

44. *Jakubiček M., Kilgariff A., Kovář V., Rychlý P., Suchomel V. The TenTen Corpus Family // Proceedings of the 7th International Corpus Linguistics Conference CL 2013, the United Kingdom, July 2013. P. 125–127 (2013).*

45. *Sinclair J. Corpus and Text – Basic Principles // Developing Linguistic Corpora: a Guide to Good Practice. Oxford : Oxbow Books, 2005. P. 1–16. URL: <http://users.ox.ac.uk/~martinw/dlc/chapter1.htm> (дата обращения: 09.07.2021).*

Attributive Collocations in the Gold Standard of Russian Collocability and Their Representation in Dictionaries and Corpora

Voprosy leksikografii – Russian Journal of Lexicography, 2021, 21, pp. 33–68. DOI: 10.17223/22274200/21/2

Maria V. Khokhlova, Saint Petersburg State University (Saint Petersburg, Russian Federation). E-mail: m.khokhlova@spbu.ru

Keywords: collocations, collocability, attributive collocations, Russian language, dictionaries, text corpora, database.

The study is supported by the Russian Science Foundation, Project No. 19-78-00091.

The article discusses how collocations are represented in Russian dictionaries and how information about them can be covered in a collocation database that is being developed. Such a resource (gold standard) can be in demand when developing applications for teaching or learning Russian as a foreign language and solving other theoretical and applied issues. The aim of the study was twofold: firstly, to analyze how ex-

planatory and specialized dictionaries of the Russian language represent collocations and hence to what extent their data coincide with each other, and, secondly, to investigate how these dictionary collocations are reflected in text corpora. This allows tracing the relation between manually collected data and modern corpora. For the study, the author used the disambiguated subcorpus and the main corpus of the Russian National Corpus (RNC) with a volume of 6 million and 321 million words, respectively, as well as the large Internet corpus ruTenTen with a volume of more than 14.5 billion words. The author considered attributive phrases built according to the “adjective/participle + noun” model. She analyzed 120 collocations with different dictionary index, i.e. the number of dictionaries in which this phrase is given. The following hypothesis was tested: high collocation frequencies correspond to the fact that the item is recorded in several dictionaries. In the analysis, nonparametric analogues of analysis of variance (Friedman and Kruskal-Wallis tests) were used to assess the statistical significance of differences in quantitative data. The frequencies of collocations in corpora of different volume and in different dictionaries were compared. In total, more than 15 thousand examples were processed, less than 0.5% of them were presented in four of the six reviewed dictionaries (five printed and one electronic). The results show data heterogeneity, items selected for a dictionary do not coincide with their frequency characteristics and thus word combinations turn out to be low-frequency. About 34% of the examples are absent in the RNC corpus with removed ambiguity, and about 12% of analyzed collocations are rare (less than 0.01 ipm) even in the ruTenTen corpus. The presence of collocations in several dictionaries indicates their higher frequencies and hence reproducibility in speech. Explanatory dictionaries and collocation dictionaries show the smallest intersection of data. The results show that the amount of data is a crucial issue, and the very phenomenon of collocability should be studied on large corpora.

References

1. Khokhlova, M.V. (2008) Eksperimental'naya proverka metodov vydeleniya kollokatsiy [Experimental verification of collocation identification methods]. *Slavica Helsingiensia*. 34. pp. 343–357.
2. Yagunova, E.V. & Pivovarova, L.M. (2010) Priroda kollokatsiy v russkom yazyke. Opyt avtomaticheskogo izvlecheniya i klassifikatsii na materiale novostnykh tekstov [The nature of collocations in Russian. Experience of automatic extraction and classification based on news texts]. *Nauchno-tehnicheskaya informatsiya. Seriya 2. Informatsionnye protsessy i sistemy*. 6. pp. 30–40.
3. Pivovarova, L., Kormacheva, D. & Kopotев, M. (2018) Evaluation of collocation extraction methods for the Russian language. In: Kopotев, M. et al. (eds) *Quantitative Approaches to the Russian Language*. London; New York: Routledge. pp. 137–157.
4. Khokhlova, M. (2018) Similarity between the Association Measures: a Case Study of Noun Phrases. In: Horák, A., Rychlý, P. & Rambousek, A. (eds) *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018*. Brno: Tribun EU. pp. 21–27.
5. Enikeeva, E. & Mitrofanova, O. (2017) Russian Collocation Extraction Based on Word Embeddings. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po mate-*

rialam ezhegodnoy mezhdunarodnoy konferentsii "Dialog" (Moskva, 31 maya – 3 iyunya 2017 g.) [Computational linguistics and intelligent technologies: Based on the materials of the annual international conference "Dialogue" (Moscow, 31 May – 3 June 2017)]. Vol. 16 (23): in 2 vols. Moscow: RSUH. [Online] Available from: <http://www.dialog-21.ru/media/3908/enikeevaevmitrofanovaoa.pdf> (Accessed: 09.07.2021).

6. Kobritsov, B.P., Lyashevskaya, O.N. & Shemanaeva, O.Yu. (2005) Snyatie leksiko-semanticheskoy omonimii v novostnykh i gazetno-zhurnal'nykh tekstakh: poverkhnostnye fil'try i statisticheskaya otsenka [Removing lexical and semantic homonymy in news and newspaper and magazine texts: Surface filters and statistical evaluation]. In: *Internet-matematika 2005. Avtomaticheskaya obrabotka veb-dannykh* [Internet mathematics 2005. Automatic processing of web data]. Moscow: Yandex. pp. 38–57.

7. Vasilisina, A.A., Zarifyan, M.S., Kazakova, P.N. & Sudarikova, E.A. Polisemiya glagolov peremeshcheniya, dvizheniya i lokalizatsii (po rezul'tatam semanticheskoy razmetki bazy dannykh Aktivnogo slovary russkogo yazyka). [Online] Available from: <http://www.dia-log-21.ru/media/3471/vasilisina.pdf> (Accessed: 09.07.2021).

8. Azarova, I.V., Sinopal'nikova, A.A. & Yavorskaya, M.V. (2004) [Principles of building the RussNet wordnet thesaurus]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computational linguistics and intellectual technologies]. Dialog 2004 Conference Proceedings. Verkhnevolzhskiy. 2–7 June 2004. Moscow: RSUH. pp. 542–547. (In Russian).

9. Bol'shakov, I.A. (2009) [KrossLeksika: A large electronic dictionary of combinations and semantic connections of Russian words]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computational linguistics and intellectual technologies]. Dialog 2009 Conference Proceedings. Vol. 8 (15). Moscow: RSUH. pp. 45–50. (In Russian).

10. Berkov, V.P. (2004) *Dvuyazychnaya leksikografiya* [Bilingual lexicography]. 2nd ed. Moscow: AST.

11. Atkins, B.T.S. & Rundell, M. (2008) *The Oxford Guide to Practical Lexicography*. Oxford: OUP.

12. Benson, M., Benson, E. & Ilson, R. (1986) *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam, Philadelphia: John Benjamins.

13. Kjellmer, G. (1994) *A dictionary of English collocations: based on the Brown corpus: in three volumes*. Oxford; New York: Clarendon Press: Oxford University Press.

14. Hanks, P. (2008) Mapping meaning onto use: a Pattern Dictionary of English Verbs. *Proceedings ACL 2008*. Utah. [Online] Available from: <https://nlp.fi.muni.cz/projects/cpa/Pattern%20Dict%20Utah.ppt> (Accessed: 09.07.2021).

15. Pezik, P. (2014) Graph-Based Analysis of Collocational Profiles. In: Jesenšek, V. & Grzybek, P. (eds) *Phraseologie im Wörterbuch und Korpus (Phraseology in Dictionaries and Corpora)*. ZORA 97. Maribor; Bielsko-Biala; Budapest; Kansas; Praha: Filozofská fakulteta. pp. 227–43.

16. Mel'chuk, I.A. et al. (1984) *Tolkovo-kombinatornyy slovar' sovremennogo russkogo yazyka. Opyty semantiko-sintaksicheskogo opisaniya russkoy leksiki* [Explanatory-combinatorial dictionary of the modern Russian language. Experiments in the

semantic and syntactic description of Russian lexis]. Vienna: Wiener Slavistischer Almanach.

17. Deribas, V.M. (1983) *Ustoychivye glagol'no-imennye slovosochetaniya russkogo yazyka* [Russian stable verb-nominal phrases]. Moscow: Russkiy yazyk.

18. Borisova, E.G. (1995) *Slovo v tekste. Slovar' kollokatsiy (ustoychivyykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov* [Word in the text. Dictionary of Russian collocations (stable combinations) with the English-Russian dictionary of key words]. Moscow: Filologiya.

19. Denisov, P.N. & Morkovkin, V.V. (2002) *Slovar' sochetaemosti slov russkogo yazyka* [Collocation dictionary of Russian words]. 3rd ed. Moscow: AST.

20. Apresyan, V.Yu. et al. (2010) *Prospekt aktivnogo slovary russkogo yazyka* [Prospectus of the active dictionary of the Russian language]. [Online] Available from: http://san.ruslang.ru/prospect_theory.pdf (Accessed: 09.07.2021).

21. *Russian National Corpus*. [Online] Available from: <http://ruscorpora.ru> (Accessed: 09.07.2021). (In Russian).

22. Biryuk, O.L., Gusev, V.Yu. & Kalinina, E.Yu. (2008) *Slovar' glagol'noy sochetaemosti nepredmetnykh imen russkogo yazyka* [Dictionary of verb collocations of non-objective Russian nouns]. [Online] Available from: http://dict.ruslang.ru/abstr_noun.php.

23. Kustova, G.I. (2008) *Slovar' russkoy idiomatiki. Sochetaniya slov so znacheniem vysokoy stepeni* [Dictionary of Russian idioms. Collocations of words with the meaning of a high degree]. [Online] Available from: <http://dict.ruslang.ru/magn.php> (Accessed: 09.07.2021).

24. Lyashevskaya, O. (2010) Bank of Russian Constructions and Valencies. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta. pp. 1802–1805.

25. Kopotev, M. et al. (2015) oCoCo: Online Extraction of Russian Multiword Expressions. *The 5th Workshop on Balto-Slavic Natural Language Processing* (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd. pp. 43–45.

26. Sharoff, S. & Nivre, J. (2011) *The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog" (Bekasovo, 25–29 maya 2011 g.)* [Computational linguistics and intelligent technologies: Based on the materials of the annual international conference "Dialogue" (Bekasovo, 25–29 May 2011)]. Vol. 10 (17). Moscow: RSUH. pp. 657–670.

27. Lexicograph Database. [Online] Available from: <http://lexicograph.ruslang.ru> (Accessed: 09.07.2021).

28. Evgen'eva, A.P. (ed.) (1999) *Slovar' russkogo yazyka: v 4 t.* [Dictionary of the Russian language: In 4 volumes]. 4th ed. Moscow: Rus. yaz.

29. Calzolari, N. et al. (2002) Towards Best Practice for Multiword Expressions in Computational Lexicons. *Proceedings of LREC*. pp. 1934–1940.

30. Lukashevich, N.V., Dobrov, B.V. & Chuyko, D.S. (2008) [The selection of word combinations for the dictionary of the automatic text processing system]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computational linguistics

and intellectual technologies]. Dialog–2008 Conference Proceedings. Moscow: RSUH. pp. 339–344. (In Russian).

31. Testeleits, Ya.G. (2001) *Vvedenie v obshchiy sintaksis* [An introduction to general syntax]. Moscow: RSUH.

32. Teliya, V.N. (1996) *Russkaya frazeologiya. Semanticheskii, pragmaticheskii i lingvokul'turologicheskii aspekty* [Russian phraseology. Semantic, pragmatic and linguocultural aspects]. Moscow: Yazyki slavyanskoy kul'tury.

33. Gorbachevich, K.S. et al. (eds) (2004–present) *Bol'shoy akademicheskii slovar' russkogo yazyka: v 20 t.* [The Great Academic Dictionary of the Russian Language: In 20 volumes]. Moscow: Institute for Linguistic Research, RAS; Nauka.

34. Kuznetsov, S.A. (ed.) (1998) *Bol'shoy tolkovyy slovar' russkogo yazyka: A–Ya* [The Great Explanatory Dictionary of the Russian Language: A–Ya]. St. Petersburg: Norint.

35. Reginina, K.V., Tyurina, G.P. & Shirokova, L.I. (1980) *Ustoychivye slovosochetaniya russkogo yazyka: ucheb. posobie dlya studentov-inostrantsev* [Stable phrases of the Russian language: Textbook for foreign students]. Moscow: Rus. yaz.

36. Ubin, I.I. (1987) *Slovar' usilitel'nykh slovosochetaniy russkogo i angliyskogo yazykov* [Dictionary of amplifying phrases of Russian and English]. Moscow: Rus. yaz.

37. Kallas, J. et al. (2019) Lexicographic Practices in Europe: Results of the ELEX-IS Survey on User Needs. *Electronic lexicography in the 21st century*. Proceedings of the eLex 2019 Conference. Sintra, Portugal. 1–3 October 2019. Brno: Lexical Computing CZ, s.r.o. pp. 519–536.

38. Vinogradov, V.V. (1972) *Russkiy yazyk* [The Russian language]. Moscow: Vysshaya shkola.

39. Shanskiy, N.M. (1985) *Frazeologiya sovremennogo russkogo yazyka: ucheb. posobie dlya stud. filol. fak-tov* [Phraseology of the modern Russian language: Textbook for philology students]. 3rd ed. Moscow: Vysshaya shkola.

40. *MyStem*. [Online] Available from: <https://yandex.ru/dev/mystem/> (Accessed: 09.07.2021).

41. MacIntosh, C. (ed.) (2009) *Oxford Collocations Dictionary for Students of English*. 2nd ed. Oxford: Oxford University Press.

42. Quasthoff, U. (2011) *Wörterbuch der Kollokationen im Deutschen*. Berlin; New York: Walter de Gruyter.

43. Lyashevskaya, O.N. & Sharov, S.A. (2009) *Chastotnyy slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)* [Frequency Dictionary of the Modern Russian Language (Based on the Materials of the Russian National Corpus)]. [Online] Available from: <http://dict.ruslang.ru/freq.php> (Accessed: 09.07.2021).

44. Jakubiček, M. et al. (2013) The TenTen Corpus Family. *Proceedings of the 7th International Corpus Linguistics Conference CL 2013*. United Kingdom. pp. 125–127.

45. Sinclair, J. (2005) Corpus and Text – Basic Principles. In: Wynne, M. (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. pp. 1–16. [Online] Available from: <http://users.ox.ac.uk/~martinw/dlc/chapter1.htm> (Accessed: 09.07.2021).