# СОЧЕТАЕМОСТЬ ЧЕРЕЗ ПРИЗМУ КОРПУСОВ

**Захаров В. П.** (vz1311@yandex.ru)

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

В статье рассматриваются устойчивые сочетания разного типа и показываются способы их количественной оценки. Описаны эксперименты, в ходе которых на материале корпусов русского языка и инструментов корпусной лингвистики показано, как с помощью корпусных методов можно расширить состав словарных статей в словарях устойчивых выражений и как можно количественно оценить употребительность и устойчивость словосочетаний в синхронии и диахронии.

**Ключевые слова:** устойчивые словосочетания, фразеологизмы, коллокации, корпусы текстов, меры ассоциации, диахронические исследования

# SET PHRASES: A VIEW THROUGH CORPORA

**Zakharov V. P.** (vz1311@yandex.ru)

Saint-Petersburg State University, Saint-Petersburg, Russia

The study of word collocatibility is one of the main tasks of linguistics. Syntagmatic relations bind together language units being in direct contact with each other. The combinatory ability of language units, collocatibility, is one of the linguistic syntagmatic laws. This phenomenon is the main object of the phraseology and lexicography. The article deals with set phrases of different types from the point of view of their numerical evaluation. Corpus linguistics understand set phrases as statistically determined unities. This approach is the basic point of different automatic ways to extract idioms and collocations. The paper describes experiments which show how text corpora and corpus methods and tools such as association measures, word sketches, concordances can be used to expand the entries in existing dictionaries and how set phrases could be evaluated quantitatively. There are a small numbers of works on set phrases productivity during time periods because of small size of historical corpora. In this research examined set phrases usage was studied diachronically on the base of the big Google books Ngram Viewer Russian corpus counting billions of tokens. The study argues that diachronic productivity is best evaluated with a studying contexts. Used corpus tools enable to do it. Ultimately, it is shown and maintained that corpus linguistics methods and tools allow to create dictionaries of new type which have to include a larger amount of set phrases and collocations than before.

**Key words:** set phrases, idioms, collocations, collocation dictionaries, corpus, association measures, concordance, diachronical research

## Введение

Один из популярных предметов в языкознании — это устойчивые словосочетания. Несмотря на пристальное внимание лингвистов к фразеологии и связанным темам, можно утверждать, что состояние фразеологии сегодня, на наш взгляд, неудовлетворительно. Фразеологический запас русского языка разбросан по разным лексикографическим изданиям, прежде всего это толковые и фразеологические словари, и ни один словарь не может считаться достаточно полным по охвату фразеологического лексикона. Предположительно, словник такого полного словаря должен насчитывать несколько сотен тысяч единиц. Также нетрудно убедиться, что статьи существующих фразеологических словарей неполны, они плохо структурированы, никак не привязаны к хронологии. В наши дни эту ситуацию можно существенно улучшить. Наличие корпусов текстов создает предпосылки для создания большого словаря сочетаемости, основанного на корпусах, с количественной параметризацией внутри.

Нужно отметить, что в корпусной лингвистике сложилась методология более широкого понимания фразеологии, и границы фразеологии здесь значительно расширены (или размыты) за счет новых подходов, общим для которых является понятие «статистической устойчивости». Может быть, самой знаменитой цитатой в корпусной лингвистике является высказывание Дж. Р. Фёрса «Вы поймете слово по его окружению» (“You shall know a word by the company it keeps”) [Firth 1957: 179]. Там же и тогда же им было введено вошедшее сегодня в широкий оборот понятие коллокации, базирующееся на статистических критериях. Об этом же писал И. А. Мельчук. «Устойчивость сочетания относительно данного элемента измеряется вероятностью, с которой данный элемент предсказывает совместное появление остальных элементов сочетания (в определенном порядке относительно предсказывающего элемента)» [Мельчук 1960: 73].

В корпусной лингвистике в основе методов вычисления силы синтагматической связи между элементами словосочетаний лежат частотные характеристики и структурно-синтаксические модели, на основе которых по формулам так называемых ассоциативных мер (мер ассоциации) вычисляется коэффициент силы связанности или, по-другому, уникальности словосочетания.

Данное исследование преследует цель показать, как можно улучшить словари сочетаемости корпусными методами. Была поставлена задача — на основе корпусных данных изучить «поведение» некоторых устойчивых словосочетаний, в том числе в течение длительного промежутка времени.

## 1. Материал и инструмент исследования

В качестве материала и инструмента исследования были использованы Национальный корпус русского языка (НКРЯ) (http://ruscorpora.ru), корпусы русских текстов ruTenTen 2011 и ruTenTen 2011 sample системы Sketch Engine (https://the.sketchengine.co.uk/), корпус русских текстов Araneum Russicum Maius из семейства псевдопараллельных корпусов Aranea Университета им. А. Коменского

в Братиславе (http://ucts.uniba.sk/) [Benko 2013], русский корпус системы Google books Ngram Viewer и, соответственно, их программные средства. Объем основного корпуса НКРЯ составляет 230 млн словоупотреблений, ruTenTen 2011 sample и русскоязычный Araneum насчитывают по 1,2 млрд токенов (около 1 млрд текстоформ), ruTenTen 2011 имеет объем более 18 млрд токенов (14,5 млрд текстоформ). Самый же большой из них — корпус русских книг Google books Ngram Viewer (https://books.google.com/ngrams). В настоящее время это наиболее мощный инструмент диахронических исследований. Эта система содержит корпусы размеченных текстов книг на 9 языках. Корпус книг на русском языке содержит 591 310 текстов общим объемом более 67 млрд словоупотреблений. Самые поздние публикации, включенные в корпус, относятся к 2008 году.

Основной лексической единицей (ЛЕ), с которой работает данная система, является N-грамма — последовательность от одной до пяти словоформ. Причем N-грамма, для того чтобы быть учтенной и обработанной, должна встретиться в корпусе не менее 40 раз. Для каждой заданной ЛЕ для заданного временного интервала строится единый график, по вертикальной оси которого откладывается относительная частота встречаемости заданных N-грамм в корпусе в данном году (частота, деленная на общее число словоупотреблений в корпусе за этот год), выраженная в процентах. На горизонтальной оси показаны годы, входящие в заданный временной интервал. *Каждая кривая графика маркируется цветом, в конце кривой указывается, какой* N-грамме (слову или словосочетанию) она соответствует (рис. 1).

При построении графиков изменения частоты употребления ЛЕ используется так называемое «сглаживание» (smoothing) При нулевом сглаживании в графике учитывается относительная частота встречаемости N-граммы за каждый год. Однако тенденция в динамике «поведения» слов более отчетливо прослеживается при скользящем усреднении данных. Если значение коэффициента сглаживания равно 3, то это означает, что для некоторого года к числу словоупотреблений искомого слова за этот год прибавляется число словоупотреблений его за три предыдущих года и три последующих и полученная сумма делится на семь. Относительное значение этой средней величины в процентах отражается на вертикальной оси.

Имеется тег «подстановочный знак» * (wildcard). Ввод его через пробел после N-граммы или до неё позволяет построить график встречаемости десяти наиболее частотных сочетаний данной N-граммы и слова, следующего за нею или ей предшествующего. Кроме построения графиков, система предоставляет ссылки к текстам, где встретились заданные ЛЕ. Как правило, это библиографические описания книг и фрагменты текстов с выделением в них заданных N-грамм. В некоторых случаях доступен полный текст книги в графическом формате. Более подробно о сервисе Google books Ngram Viewer см. [Захаров, Масевич 2014].

Похожий инструмент под названием «Графики» с 2012 г. работает и в составе НКРЯ. Функционально он подобен сервису Google books Ngram Viewer. Вход в этот сервис возможен как со страницы с результатами поиска по запросу к основному корпусу (ссылка *Распределение по годам*), так и из главного меню (ссылка *Графики*). При сходной идеологии, формулы подсчета относительной частоты и сглаживания в сервисах Национального корпуса и Google Ngram Viewer отличаются. Имеется возможность показать таблицы с абсолютными

и относительными частотами заданных ЛЕ за каждый год. Из таблиц по гиперссылкам возможен переход к просмотру примеров из корпуса.

## 2. Эксперименты

В качестве примеров устойчивых сочетаний для исследования были выбраны сочетания двух типов: 1) свободные сочетания с характерными определениями к слову «аплодисменты», выражающими, говоря в терминах теории «Смысл — Текст», функцию Magn; 2) фразеологизированные сочетания с глаголом «перебиваться» в значении «бедствовать».

### 2.1. «Аплодисменты»

Посмотрим, какие стандартные определения к слову «аплодисменты» зафиксированы в словарях. Новый Большой академический словарь приводит следующие сочетания: *бурные аплодисменты*, *гром аплодисментов* [БАС 2004]. Словарь сочетаемости слов русского языка дает: *громкие, продолжительные, долго не смолкающие, несмолкаемые, бурные, дружные, одобрительные, горячие, восторженные аплодисменты,* а также *сдержанные, скупые, редкие, жидкие* [Словарь сочетаемости 1983]. Неизвестно, что отражает порядок их следования.

В системе Sketch Engine имеется инструмент вычисления коллокаций по 7 мерам ассоциации. В одном из режимов мы получили список из 36 прилагательных, из которых 19 можно считать функцией Magn от слова «аплодисменты». Вот список этих прилагательных, упорядоченный по алфавиту: *бешеный, бурный, дружный, восторженный, всеобщий, горячий, громкий, несмолкающий, громовой, громогласный, долгий, дружный, неистовый, нескончаемый, несмолкаемый, несмолкающий, оглушительный, продолжительный, шумный.*

Традиционные словари ничего не говорят о частоте употребления словарных единиц. Эти данные можно получить из корпусов. Например, поиск в НКРЯ (интервал 3 слова вправо) в данном случае дает следующие цифры: *бурные аплодисменты* 337 словоупотреблений, *продолжительные* 125. *дружные* 81, *громкие* 47, *оглушительные* 22, в то время как *несмолкающие* всего одно.

При этом важно понимать, в каком корпусе мы ищем, что и как ищем. Так, при поиске словосочетания *бурные аплодисменты* в НКРЯ в интервале в одно слово вправо оно было найдено в 279 контекстах, в то время как поиск в интервале в 3 слова дает нам 337 контекстов. Подавляющая часть прироста обеспечивается релевантным сочетанием *бурные и продолжительные аплодисменты*. Поиск же по сочетанию *дружные аплодисменты* в том же интервале выдает нам не совсем корректные сочетания *дружный смех и аплодисменты*, *дружный хохот и аплодисменты* и др. То есть, число 81 для *дружных аплодисментов* несколько завышено. Главное, полученные цифры не нужно абсолютизировать, важны их относительные величины.

Иногда, задав нестандартный режим поиска, можно получить дополнительно интересные результаты. Например, ни один из наших корпусов не дал

к *аплодисментам* коллоката *жесткий*. Однако поиск в интервале с отключенным согласованием между этими словами дает фразу *«По жесткому звуку аплодисментов чувствовалось…»*, где появляется это определение.

Выявление коллокаций по формулам мер ассоциации учитывает не только частоту совместной встречаемости, но и частость или редкость каждого элемента, тем самым мы вычисляем именно силу связи между элементами сочетания. Результат можно упорядочить или по частоте, или по значению меры ассоциации (табл. 1). И мы видим, что *продолжительные аплодисменты* по силе связи (мера salience) оказались лишь на 7-м месте.

**Таблица 1.** Примеры сочетаний «прилагательное + *аплодисменты*» в корпусе ruTenTen 2011, упорядоченных по мере ассоциации salience

| № п/п | Словосочетание | Частота в корпусе | Мера salience |
|---|---|---|---|
| 1. | бурные аплодисменты | 13 372 | 10,25 |
| 2. | дружные аплодисменты | 2 051 | 8,42 |
| 3. | оглушительные аплодисменты | 656 | 8,29 |
| 4. | громкие аплодисменты | 3 711 | 8,22 |
| 5. | восторженные аплодисменты | 899 | 8,17 |
| 6. | одобрительные аплодисменты | 388 | 8,05 |
| 7. | продолжительные аплодисменты | 2 495 | 7,80 |
| 8. | несмолкающие аплодисменты | 211 | 7,62 |

Результаты поиска сочетаний со словом *аплодисменты*, полученные разработчиками Генерального Интернет-корпуса русского языка (ГИКРЯ) [Беликов и др. 2013] на их корпусе (подкорпус «Журнальный зал», объем 313 млн словоупотреблений) и любезно предоставленные автору, дают несколько другую картину, а именно: *бурные* 194, *продолжительные* 72, в т. ч. *бурные (и) продолжительные* 33, *громкие* 39, *дружные* 26, *шумные* 17, *восторженные* 15, *долгие* 14, *горячие* 13, *оглушительные* 12, *несмолкающие* 12. Это еще раз говорит о том, что цифры не нужно абсолютизировать и нужно учитывать, на каком корпусе они получены. При этом для оценки распространенности соответствующих единиц в корпусе (а тем самым в какой-то степени и в языке) нужно опираться не на абсолютные частоты, а на относительные (ipm). Проиллюстрируем это на маленьком примере (табл. 2).

**Таблица 2.** Сравнение частот сочетаний в разных корпусах

| | Частота в корпусе | | | ipm | | |
|---|---|---|---|---|---|---|
| Словосочетание | НКРЯ | ruTenTen | ГИКРЯ | НКРЯ | ruTenTen | ГИКРЯ |
| дружные аплодисменты | 81 | 2 051 | 26 | **0,350** | 0,140 | 0,080 |
| громкие аплодисменты | 47 | 3 711 | 39 | 0,200 | **0,260** | 0,120 |
| несмолкающие аплодисменты | 1 | 211 | 12 | 0,004 | 0,015 | **0,038** |

Как видим, разные корпусы по-разному оценивают вес соответствующих сочетаний в языке, а фактически, в подъязыке, который представлен корпусом. Отдельная задача — попытаться эту разницу понять и объяснить, с тем чтобы какие-то особенности корпуса (преобладание какой-то тематики или типа текстов, возможное наличие дублетов и т.п.) не переносить на язык в целом.

Тем не менее, создавая словарь устойчивых сочетаний на основе корпусов, мы имеем возможность выстроить их по частоте употребления или по силе «спаянности», оговорив, на каком материале этот словарь создается. Более того, по-видимому, иногда полезно опираться на усредненные характеристики, полученные на разных корпусах.

Данные, полученные на синхронных корпусах, ничего не говорят о продуктивности ЛЕ в разные промежутки времени. Чтобы увидеть их использование на протяжении длительного периода, построим графики распределения частоты употребления наших сочетаний в сервисах Google books Ngram Viewer и «Графики» НКРЯ в текстах двух последних столетий (рис. 1).


**Рис. 1.** Кривые встречаемости биграмм со вторым словом «аплодисменты» в корпусе Google books Ngram Viewer (сглаживание 3)

Как мы уже видели, «лидируют» *бурные аплодисменты*, а на втором месте — *продолжительные*. Но это суммарные данные по всему корпусу. На графиках же мы видим особенности распределения частот употребления этих сочетаний во времени — см. пики на рубеже 1940-х, 1960-х и 1980-х годов. И если во второй половине 1930-х «верх берут» *бурные*, то в конце 70-х — начале 80-х преобладают *продолжительные*. Следует отметить, что в широкое употребление все эти сочетания вошли только в XX веке. А в конце века их частотность резко упала. Это видно и из сервиса Google, и сервиса НКРЯ. Однако сервис НКРЯ в этом и в других случаях дает картину мало репрезентативную по причине недостаточности данных. Анализируемые словосочетания представлены в корпусе в малых количествах и не в каждом году. Так, *громкие аплодисменты* встретились в основном корпусе НКРЯ по одному разу в 1885, 1906, 1908, 1910, 1925, 1939–40, 1959, 1963, 1998–2000, 2003 гг. и два — в 2001 г. Этого явно мало. Поэтому мы опираемся на графики системы Google.

Если же задать сглаживание, равное нулю, то можно определить пик использования того или другого сочетания в каждом году (точнее, в текстах

данного года; напомним, в корпусе Google books Ngram Viewer это книги). Для *продолжительных аплодисментов* «рекордным» оказался 1981 г.

Проанализируем также атрибутивное отношение, выраженное в форме «существительное в им. пад. + *аплодисменты* в род. пад.». Все словари согласно приводят следующие коллокаты для выражений этого типа: *шквал, гром, буря, грохот, взрыв*. Остается, однако, неясной частотность их употребления. Данные поиска в корпусах ruTenTen 2011 sample и Araneum Russicum Maius и график (рис. 2) показывают явное преимущество коллоката *гром*.


**Рис. 2.** Кривые встречаемости биграмм с существительным и со вторым словом «аплодисменты» в родительном падеже в корпусе Google books Ngram Viewer (сглаживание 3)

На втором месте — *буря*. Но мы видим, что начиная с 1980-х годов *шквал аплодисментов* идет вверх и устойчиво обгоняет *бурю*. А если обратиться к корпусу ГИКРЯ, отражающему современное состояние языка, то в подкорпусах «Живой журнал» и «В контакте» *шквал* уже обошел и *гром*, что говорит о том, что «живой» язык, видимо, предпочитает последнее сочетание.

Поиск в основном корпусе НКРЯ дает 118 вхождений для *гром аплодисментов* (165 с учетом словоизменения, но следует помнить, что графический сервис работает со словоформами), 33 вхождения для *шквал аплодисментов*, 15 для *буря аплодисментов*, только одно для *взрыв аплодисментов* и ни одного для *грохота*.

Однако можем ли мы полностью доверяться конкретному корпусу? Последние два сочетания в 230-миллионном НКРЯ фактически не представлены, зато в соизмеримом с НКРЯ по объему подкорпусе ГИКРЯ («Журнальный зал») нашлось 25 *взрывов* и 10 *грохотов*. Этот и подобные факты требуют своего объяснения, чтобы мы могли опираться на получаемые результаты либо иногда их отбрасывать. Например, частотное слово или словосочетание может иметь источником этой частоты его «взрывообразную» представленность всего лишь в нескольких текстах в коротком промежутке времени и не быть характерным для языка в целом. Для минимизации таких «всплесков» в лингвистике существуют специальные меры, учитывающие равномерность появления слова в корпусе (коэффициент Жуйана, Average Reduced Frequency и др.).

И последнее: следует учитывать лексико-синтаксическое варьирование исследуемых сочетаний. Так, в корпусе Araneum Russicum Maius корпусный менеджер насчитал 14 *взрывов аплодисментов*, в то время как сочетание *взорваться аплодисментами* в разных формах глагола встретилось 42 раза, что говорит о том, что набор конструкций, выражающих функцию Magn от слова «аплодисменты», должен быть расширен.

## 2.2. «Перебиваться»

Многие фразеологизмы и устойчивые сочетания имеют лексико-синтаксические варианты, когда либо меняется лексическое наполнение в рамках некоторой структурной формулы, либо при том же наполнении меняется формула. Например, «кошки скребут». Но где? Словари сообщают, что *на душе* и *на сердце*. А где чаще? По данным корпуса Google books Ngram Viewer выясняется, что чаще *кошки скребут на душе* и больше всего они скребли в годы на переломе 1980–90-х гг.

Наверное, не будет ошибкой утверждение, что фразеологизмов с лексико-синтаксическими ввариациями большинство. Примеры их можно множить и множить: *беречь (хранить) как зеницу ока; беречь пуще глаза; мерить одной мерой (меркой), мерить на одну меру (мерку); ест за троих, есть в три горла; драть (сдирать/содрать) шкуру (три, две шкуры); драть (сдирать/содрать) по три (две) шкуры;хоть в землю заройся, хоть из-под земли достань; брать/взять (забирать/забрать) в [свои] руки, прибирать/прибрать к рукам; сталкивать/столкнуться лицом к лицу, носом к носу, нос в нос, лоб в лоб* [Бирих и др. 1997]. Такие вариативные сочетания в словарях описаны, естественно, менее полно по сравнению с лексикализованными фраземами.

Рассмотрим этот тип вариаций более подробно на примере сочетания глагода «перебиваться» с предложной конструкцией «с … на …». Традиционно словари дают сочетания *перебиваться с хлеба на квас, перебиваться с хлеба на воду*. Кроме того, приводятся синонимические конструкции *с куска на кусок, с гроша на копейку, с пуговки на петельку* [Бирих и др. 1997: 15].

Посмотрим, что нам дополнительно дают корпусы. Поиск в указанных корпусах добавляет к вышеприведенному списку еще немалую толику, а именно: *с хлеба на воду,* с *хлеба на кофе, с гроша на грош, с копейки на копейку, с рубля на рубль, с хлеба на квас, с воды на квас, с воды на хлеб, с хлеба на картошку, с петельки на пуговку, со дня на день, с весны на весну, с работы на работу.* Есть и более экзотические: *с селедки на вермишель, с "Российского" на "Докторскую"*. А к вершинам народного языкового творчества можно отнести вот это: «И это беспроводной интернет!.. У Мегафона четкий прием — 3894 Кбит/с. Мне, молившемуся на dial-up, который *перебивался с 22 на 40 килобит в секунду*, это кажется чем-то фантастическим».

Вышеприведенные примеры в основной массе получены из корпусов, созданных по технологии Wacky, то есть на основе текстов из веба, и нередко они отражают языковое творчество, но не узус, подчеркивая лишь продуктивность

конструкции «перебиваться с … на …». Однако наличие больших корпусов позволяет выявлять значительное число кандидатов на вхождение в словарь в качестве устойчивых сочетаний, а статистические данные дают возможность оценить распространенность того или другого сочетания.

Из графика на рис. 3 видно, что наиболее частые устойчивые сочетания с глаголом «перебиваться» это те, которые приводятся в фразеологических словарях, и что активно в языковой обиход они вошли только в начале XX века. Зато сочетание *перебиваться со дня на день* широко использовалось в XIX веке.

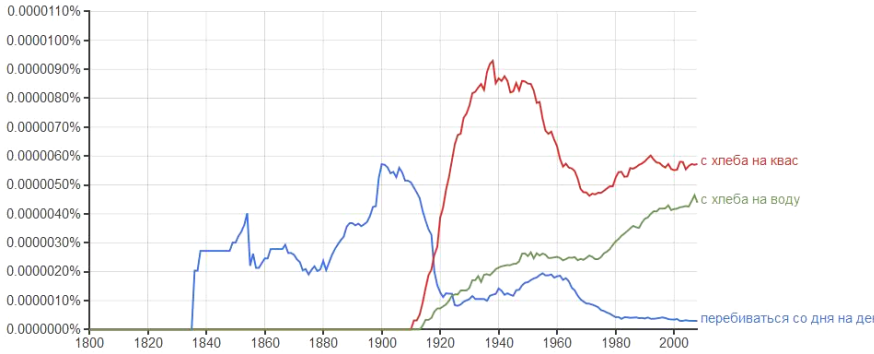


**Рис. 3.** Кривые встречаемости выражений с глаголом «перебиваться» в корпусе Google books Ngram Viewer (сглаживание 3)

И здесь еще раз следует подчеркнуть, что, интерпретируя корпусные данные, мы должны хорошо понимать, что собой представляет тот или другой корпус и как эти данные получены. Например, данные анализа по корпусу ГИКРЯ, предоставленные автору его разработчиками, показывают, что если в корпусе книг Google сочетание *с хлеба на квас* встречается чаще, чем *с хлеба на воду*, то во всех трех подкорпусах ГИКРЯ картина диаметрально противоположная: в сумме 269 употреблений *с хлеба на воду* против 97 *с хлеба на квас*. То же соотношение демонстрирует и корпус ruTenTen 2011 (787 против 370). Все это позволяет говорить о различии в использовании этих выражений между книжным языком и современным «спонтанным».

## 3. Опыт корпусного исследования

Проведем небольшое исследование на тему, как выглядит в корпусах сочетаемость слова «мастер». Сочетаемость слов определяется различными факторами: лексическими, грамматическими, семантическими, стилистическими. Все они влияют на норму и на узус. Можно утверждать, что узус — один из определяющих факторов при составлении словарей. И один из подходов к изучению узуса заключается в выявлении статистических закономерностей на корпусах текстов.

Слово «мастер» в русском языке является довольно частотным. Его ipm по электронной версии Частотного словаря современного русского языка (на материалах Национального корпуса русского языка) О. Н. Ляшевской и С. А. Шарова (http://dict.ruslang.ru/freq.php) равняется 100,8, в корпусе ruTenTen 2011 он равен 96,0. В словарях его сочетаемостные свойства никак особенно не описываются.

Рассмотрим его «поведение» на материале вышеупомянутых корпусов. При изучении конкордансов с этим словом обращаешь внимание на сочетание «каких-то дел мастер». И таких контекстов, кроме обычно приводимого в словарях *золотых дел мастер*, в корпусах находится достаточно много (329 в НКРЯ и 4015 в ruTenTen 2011). Если объединить все определения к словоформе *дел*, то их будет 252:

*абордажных, автомобильных, аккордеонных, алмазных, багетных, балаганных, банкетных, банных, барабанных, баррикадных, бархатного, берестяных, библиотечных, бриллиантовых, броневых, бронзовых, бронзовых, бронно-кольчужных, булочных, бумажных, буровых, бытовых, взрывных, винных, витражных, водочных, воровских, выборных, вывесочных, газетных, газовых, гаражных, гармонных, гитарных, глазных, гламурных, глиняных, гончарных, горных, городовых, грильных, гробовых, дамских, дверных, деревянных, деспотических, дипломатических, добрых, домашних, дорожных, железных, жестяных, живописного, журнальных, закулисных, замочных, заплатных, заплечных, запретных, здоровых, земельных, зеркальных, золотых, игрушечных, изразцовых, именных, иностранных, искусных, кабельных, кальянных, каменных, каминных, каретных, карманных, картежных, кирпичных, ключных, книжных, ковровых, кожаных, кожевенных, колбасных, колесных, колодезных, колокольных, колыбельных, кольчужных, комедийных, компьютерных, конкурсных, конфетных, коньячных, копировальных, корабельных, корабельных, костяных, кофейных, красочных, крепостных, крепостных, кроватных, кровельных, кровопийственных, кузнечных, кузовных, кукольных, кулачных, кулинарных, кухонных, ледяных, лепных, литейных, литературных, лодочных, любовных, макетных, малярно-живописных, малярных, машинных, мебельных, мебельных, медных, мироедских, мозаичного, мозольных, молочных, монетных, мостовых, музыкальных, музыкальных, мусийных, мясных, ножевых, обувных, огненных, оконных, оловянных, оптических, органных, оружейных, открыточных, палатных, палаточных, палаческих, памятных, парусных, переговорных, переплетных, персонных, перспективных, перчаточных, песочных, печатных, печных, плотницких, погребальных, поддельных, подкопных, подручных, подъемных, позолотных, половых, помойных, портновских, портных, портняжных, похоронных, почтовых, преоспективного, прикладных, пробирных, прохвостных, пушечных, пыточных, пытошных, ракетных, резных, рекламных, ресторанных, ритуальных, розыскных, ручных, рыбных, садовых, самолетных, сапожных, сателлитных, седельных, селедочных, сердечных, серебряных, сих, скрипичных, сладких, слесарных, словесных, социальных, ссудных, стегательных, стеклянных, стекольных, стекольных, столярных, страховых, строительных, суконных, сусалнаго, сценических, сыскных, табачных, табуреточных, тайных,*

*телевизионных, ткацких, токарных, топорных, трубных, угольных, ударных, фальшивых, фейерверкского, фершельных, фискальных, фонтанных, фотографических, фотошопных, хлебных, ходульных, холодильных, хрустальных, цветочных, ценинных, цеховых, циркульных, чайных, часовых, чеканного, чемоданных, черепаховых, чернильных, шапочных, шашлычных, швейных, шлифовальных, шлюзных, шляпных, шляпочных, шоколадных, ювелирных, янтарных.*

Анализ показывает, что за исключением немногих, окказиональных или оценочных, все они относятся к какому-либо ремеслу. Приведем 25 наиболее частотных сочетаний, полученных на корпусе ruTenTen (табл. 3).

**Таблица 3.** Наиболее частотные прилагательные в сочетаниях типа «таких-то дел мастер», упорядоченные по частоте совместной встречаемости

| Ранг по частоте | Слово | Частота |
|---|---|---|
| *1.* | золотых | 519 |
| *2.* | часовых | 261 |
| *3.* | серебряных | 179 |
| *4.* | заплечных | 112 |
| *5.* | каменных | 111 |
| *6.* | кукольных | 92 |
| *7.* | гробовых | 73 |
| *8.* | пыточных | 58 |
| *9.* | витражных | 57 |
| *10.* | добрых | 49 |
| *11.* | оружейных | 46 |
| *12.* | кузнечных | 46 |
| *13.* | кузнечных | 46 |

| Ранг по частоте | Слово | Частота |
|---|---|---|
| *14.* | колокольных | 44 |
| *15.* | ювелирных | 40 |
| *16.* | шляпных | 39 |
| *17.* | чемоданных | 38 |
| *18.* | обувных | 38 |
| *19.* | похоронных | 37 |
| *20.* | мебельных | 37 |
| *21.* | деревянных | 36 |
| *22.* | сапожных | 33 |
| *23.* | печатных | 31 |
| *24.* | скрипичных | 30 |
| *25.* | столярных | 27 |

Высокая величина частоты совместной встречаемости, казалось бы, говорит об устойчивости данного сочетания. Однако этой характеристики недостаточно, чтобы говорить о предпочтительной сочетаемости одного слова с другим. Сочетание с невысокой частотой совместной встречаемости может представлять собой неделимое единство. Имеется целый ряд статистических мер (меры ассоциации, англ. association measures), вычисляющих силу «спаянности» сочетаний. Значения мер ассоциации можно считать показателями силы синтагматической связи между элементами словосочетаний. Приведем те же 25 сочетаний с подсчитанными значениями нескольких мер ассоциации (табл. 4).

Мы видим, что коллокации «перестроились» и что разные меры по-разному оценивают силу синтагматической связи. И не всегда бóльшую силу связи получают наиболее частые сочетания, например, сочетание *витражных дел мастер*, всего лишь девятое по частоте (табл. 4), оказывается первым по рангу меры MI3 и, по-видимому. с бóльшим основанием может быть включено в словарь в качестве (или как пример) устойчивого словосочетания.

**Таблица 4.** Наиболее частотные прилагательные в сочетаниях типа «таких-то дел мастер», упорядоченные по значению меры ассоциации MI3

| Ранг по частоте | Слово | Частота сочетания | Ранг по MI3 | MI3 | log likelihood | log Dice | MI. log_f |
|---|---|---|---|---|---|---|---|
| *9.* | витражных | 57 | *1.* | 33,096 | 1472,831 | 8,651 | 85,745 |
| *1.* | золотых | 519 | *2.* | 32,046 | 9180,223 | 6,822 | 87,557 |
| *4.* | заплечных | 112 | *3.* | 31,295 | 2132,625 | 9,131 | 82,657 |
| *7.* | гробовых | 73 | *4.* | 30,744 | 1421,884 | 8,719 | 77,587 |
| *2.* | часовых | 261 | *5.* | 30,188 | 4034,095 | 7,368 | 78,951 |
| *17.* | чемоданных | 38 | *6.* | 29,400 | 822,664 | 7,993 | 68,098 |
| *13.* | кузнечных | 46 | *7.* | 28,675 | 520,601 | 7,321 | 60,985 |
| *20.* | мебельных | 37 | *8.* | 28,365 | 594,114 | 7,550 | 61,788 |
| *16.* | шляпных | 39 | *9.* | 28,330 | 771,052 | 7,909 | 64,324 |
| *3.* | серебряных | 179 | *10.* | 28,301 | 2676,989 | 6,488 | 69,416 |
| *8.* | пыточных | 58 | *11.* | 28,261 | 971,907 | 8,109 | 66,014 |
| *6.* | кукольных | 92 | *12.* | 27,395 | 1239,558 | 7,604 | 64,266 |
| *14.* | колокольных | 44 | *13.* | 27,085 | 799,089 | 7,754 | 60,775 |
| *23.* | печатных | 31 | *14.* | 27,023 | 584,485 | 7,532 | 58,078 |
| *22.* | сапожных | 33 | *15.* | 26,760 | 556,710 | 7,464 | 56,926 |
| *24.* | скрипичных | 30 | *16.* | 25,642 | 534,014 | 7,264 | 53,529 |
| *5.* | каменных | 111 | *17.* | 25,584 | 1529,471 | 5,062 | 56,727 |
| *10.* | добрых | 49 | *18.* | 25,089 | 650,132 | 6,918 | 53,147 |
| *11.* | оружейных | 46 | *19.* | 25,050 | 719,464 | 6,724 | 53,500 |
| *21.* | деревянных | 36 | *20.* | 25,019 | 414,504 | 7,010 | 49,856 |
| *12.* | кузнечных | 46 | *21.* | 24,712 | 500,269 | 6,942 | 50,463 |
| *19.* | похоронных | 37 | *22.* | 24,203 | 510,628 | 6,642 | 49,124 |
| *18.* | обувных | 38 | *23.* | 23,816 | 509,175 | 6,364 | 47,956 |
| *25.* | столярных | 27 | *24.* | 22,038 | 344,005 | 5,640 | 40,968 |
| *15.* | ювелирных | 40 | *25.* | 20,973 | 423,087 | 3,777 | 38,368 |

В лексикографических изданиях следует, вероятно, также указать, что в этих сочетаниях слово *дело* почти всегда стоит во множественном числе (из 329 сочетаний в НКРЯ только в 17 *дело* в единственном числе). Также нужно упомянуть, что в этого рода сочетаниях имеет место именно такой порядок слов. Конечно, не будет большой ошибкой сказать, *мастер кузнечных дел*, но говорят ли так? Проверка на большом «живом» материале дает нам ответ: говорят, но редко. Вот что показывает корпус ruTenTen 2011: *мастер золотых дел* 12 сочетаний против 579 с *мастером* в постпозиции, *мастер серебряных дел* 12 против 179, *мастер гробовых дел* 2 и 73, *мастер пыточных дел* 1 и 58.

Стоит упомянуть в словарях и характерное сочетание *сих дел мастер*. В НКРЯ оно встречается 4 раза, в корпусе ruTenTen 2011 12, но и во всем вебе в Яндексе — всего 722 раза, причем это все дубли, а разных цитат немногих более 12.

Наиболее часто представлено высказывание Л. Троцкого о В. Ленине — о склоке, «которую разжигает *сих дел мастер* Ленин, этот профессиональный эксплуататор всякой отсталости в русском рабочем движении». Изредка это сочетание встречается в литературе, в частности, у Н. К. Михайловского, А. П. Чехова, Н. А. Тэффи, С. В. Максимова, Н. Е. Врангеля, И. Н. Потапенко, В. Ф. Пановой, причем нередко в кавычках. На самом деле же это сочетание того же профессионально-ремесленного происхождения, что и вышеприведенные выражения. Вот откуда это пошло, читаем: *«Как в Киеве я смеялся, смотря на вывеску, на которой были изображены самовар, мельница и ножницы и подписано: «сих дел мастер»…»*.

Также для анализа сочетаемости представляют интерес предложные сочетания со словом *мастер* (предлоги «по», «на», «с», «в», «от»), где оно выступает как «хозяин» в структуре зависимостей, то есть управляет предлогом, связывающим его со знаменательным словом.

Рассмотрим здесь сочетания только с предлогом «на». Очень часто это синтаксема (отношение между хозяином, предлогом и слугой), которую Г. А. Золотова определяет как потенсив — «синтаксема от отвлеченных имен, обозначающих потенциальное действие при словах модальной семантики (глаголах, именах, прилагательных). С личными именами (мастер, мастак, охотник) потенсив образует сочетание, представляющее собой модальную и экспрессивно-оценочную модификацию предикативной характеристики лица» [Золотова 1988: 197]. Отметим, что для таких сочетаний имеется синонимичная конструкция с глаголом (*мастер на шутки — мастер шутить*).

Какие же «процессные существительные» встречаются в корпусах в сочетании с *мастером*? В корпусе ruTenTen 2011 находится 34 таких сочетания со словом *дело*, причем *дело* всегда стоит во множественном числе и с определением. Частые определения к *делам эти* и *такие*, кроме того, встретились *темные, плохие, маленькие* и *пытошные*. В числе других «потенциальных действий» для слова *мастер* были найдены: *штуки, штучки, шутки, проделки, операции, авантюры, интриги, трюки, разговоры, анекдоты, выдумки, флешмобы, проказы, хитрости, слова*. Почти всегда с определениями, среди которых преобладают *такие, всякие, подобные, разные, всевозможные*, также встречаются *сходственные, веселые, подлые, хаккерские, жестокие, недобрые*. В 15 случаях следом за предлогом идут сочетания *всякого, такого, разного, различного рода*.

Встречаются в корпусах и фразеологизированные выражения:

- *мастер на все руки* (78 раз в НКРЯ, 2910 раз в ruTenTen 2011);
- *мастер с большой буквы* (4 вхождения в НКРЯ и 470 в ruTenTen 2011, еще 7 вхождений в ruTenTen с другими определениями к *букве*: *самая большая, высокая, огромная, маленькая, та самая, вышитая*);
- *дело мастера боится* (30 вхождений в НКРЯ, 260 в ruTenTen 2011, с определениями *всякое, любое, ночное дело*).

Интересны сочетания *мастера* с выражениями на *свой лад, вкус, глаз* — в этом случае *мастеру* всегда предшествуют определения *всякий, всяк, каждый*.

Есть и другие сочетания для слова *мастер* и с другими предлогами, но на этом мы здесь остановимся.

## 4. Заключение и выводы

Сегодня русский язык переживает период быстрого обновления своего состава. Не избежали этого и устойчивые сочетания. На периферии языка оказываются сочетания, отражающие некоторые стороны социальной жизни дореволюционного общества (мир чиновничества, картежные игры и др.). Зато повышенную частотность получают сочетания из области науки, техники, спорта. Чтобы увидеть все эти изменения, нужны большие корпусы, особенно для фразеологии, учитывая сравнительно низкую частоту употребления фразеологизмов в текстах. И сейчас такие корпусы начинают появляться.

В то же время необходимо, чтобы корпусная лингвистика развивала свои средства. Так, система Google Books Ngram Viewer предоставляет большие возможности для историко-культурных и лингвистических исследований. Однако в текстах корпуса встречается много ошибок распознавания. Поиск заданных лексических единиц ведется по словоформам, а не по леммам. Корпус построен исключительно на книгах и тем самым не сбалансирован. По-видимому, целесообразно было бы провести основательное исследование с применением методов статистической обработки данных, чтобы понять, как эти и другие проблемы влияют на достоверность получаемых результатов. Все это относится и к сервису НКРЯ «Графики» (главный недостаток малый для полноценных диахронических исследований объем корпуса).

Проведенное исследование показало, что корпусы и инструментарий корпусной лингвистики позволяют выявить и существенно расширить лексический фонд устойчивых словосочетаний разного типа и особенности их бытования. Основываясь на корпусах, лингвисты имеют возможность создавать словари и учебники нового типа, где сочетаемость будет представлена неизмеримо шире, чем до сих пор. В качестве примера такого словаря можно привести словарь «КроссЛексика», в котором словосочетания составляют самую важную и самую объемную его часть (2,26 млн словосочетаний) [Большаков 2009]. При этом они должны иметь количественные характеристики как силы устойчивости в синхронии, так и историю их употребления в диахронии.

В ходе исследования мы также неоднократно убеждались, что для того, чтобы можно было делать достоверные выводы на основе корпусных данных, следует хорошо представлять себе недостатки и ограничения тех инструментов, которыми мы пользуемся.

## Литература

1. *БАС* — Большой академический словарь русского языка. Том 1. М.—СПб.: Наука, 2004.
2. *Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А.* Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). М.: Изд-во РГГУ, 2013. С. 84–95.

3. *Бирих А. К., Мокиенко В. М., Степанова Л. И.* Словарь фразеологических синонимов русского языка. Ростов-на-Дону, 1997.
4. *Большаков И. А.* КроссЛексика — большой электронный словарь сочетаний и смысловых связей русских слов // Компьютерная лингвистика и интеллектуальные технологии. Международная конференция «Диалог 2009». Вып. 8 (15) М.: Изд-во РГГУ, 2009. С. 45–50.
5. *Захаров В. П., Масевич А. Ц.* Диахронические исследования на основе корпуса русских текстов Google books Ngram Viewer // Структурная и прикладная лингвистика. Вып. 10. СПб.: Изд-во С.-Петербю ун-та, 2014. С. 303–327.
6. *Золотова Г. А.* Синтаксический словарь. М.: Наука, 1988.
7. *Мельчук И. А.* О терминах «устойчивость» и «идиоматичность» // Вопросы языкознания. 1960, № 4. С. 73–80.
8. *Словарь сочетаемости слов русского языка* / Институт русского языка им. А. С. Пушкина; Под ред. П. Н. Денисова, В. В. Морковкина. — 2-е изд., испр. — М.: Русский язык, 1983.
9. *Benko V. Aranea: Yet Another Family of* (Comparable) Web Corpora. In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014, pp. 257–264. ISBN: 978-3-319-10815-5.
10. *Firth, J. R.* 1957. A synopsis of linguistic theory 1930–1955. In: F. Palmer (Ed.), Selected Papers of J. R. Firth 1952–1959. London: Longman, pp. 168–205.

## References

1. *BAS* (2004), Great Academic Dictionary of the Russian Language, [Bol'shoy akademicheskiy slovar' russkogo yazyka], vol. 1, MoscowSaint-Petersburg, Nauka.
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation, [Korpus kak yazyk: ot masshtabiruyemosti k differentsial'noy polnote], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii "Dialog 2013"], vol. 12 (19), Moscow, RGGU, pp. 84–95.
3. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora, In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264, ISBN: 978-3-319-10815-5.
4. *Birikh A. K., Mokiyenko V. M., Stepanova L. I.* (1997), Dictionary of phraseological synonyms of the Russian language, [Slovar' frazeologicheskikh sinonimov russkogo yazyka], Rostov-on-Don.
5. *Bolshakov I. A.* (2009), CrossLexica: a large electronic dictionary of collocations and semantic links between Russian words, [KrossLexika — bol'shoy

elektronnyy slovar' sochetaniy i smyslovykh svyazey russkikh slov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoj konferentsii "Dialog 2009"], vol. 8 (15), Moscow, RGGU, pp. 45–50.

6. *Collocability Dictionary of Russian Language Words* (1983), [Slovar' sochetayemosti slov russkogo yazyka], P. N. Denisov, V. V. Morkovkin (eds.), Moscow, Russkiy yazyk.
7. *Firth, J. R.* (1957), A synopsis of linguistic theory 1930–1955, In: F. Palmer (Ed.), Selected Papers of J. R. Firth 1952–1959, London, Longman, pp. 168–205.
8. *Melčuk I. A.* (1960), About the terms steadiness and idiomaticity, [O terminakh ‚ustoyvhivost" i ‚idiomatichnost"], Questions of Linguistics, [Voprosy yazykoznaniya], 1960, No. 4, pp. 73–80.
9. *Zakharov V. P., Masevich A. Ts.* (2014), Diachronic researches on the base of the Russian Google books Ngram Viewer text corpus [Diakhronicheskiye issledovaniya na osnove korpusa russkikh tekstov Google books Ngram Viewer], Srtuctural and Applied Linguistics [Strukturnaya i prikladnaya lingvistika], vol. 10, Saint-Petersburg, pp. 303–327.
10. *Zolotova G. A.* (1988), Syntactic dictionary [Sintaksicheskiy slovar'], Moscow, Nauka.