***М. В. Хохлова, И. Д. Мамаев***

**РАЗРАБОТКА БАЗЫ ДАННЫХ КОЛЛОКАЦИЙ:
ОБЗОР ЗОЛОТОГО СТАНДАРТА
НА ПРИМЕРЕ АТРИБУТИВНЫХ СЛОВОСОЧЕТАНИЙ[[1]](#footnote-1)**

**Аннотация.** Исследование лексической сочетаемости представляет собой важную проблему современной лингвистики. Статья посвящена разработке базы данных коллокаций для русского языка, которая включает словосочетания из словарей, а также примеры из корпусов текстов, которые были получены при помощи автоматических методов. Описываются основные принципы, которые лежат в основе данного ресурса, на примере отображения информации для атрибутивных словосочетаний, которые были извлечены из ряда словарей русского языка.

**Ключевые слова.** Коллокации, база данных, словари, статистические методы, русский язык

**1. Введение**

Исследование лексической сочетаемости неразрывно связано с корпусами текстов, которые дают возможность изучить примеры на большом материале. Статистические методы позволяют автоматически извлечь словосочетания на основе размеченных коллекций текстов. Однако, несмотря на то, что в настоящий момент создаются разнообразные онлайн-ресурсы, традиционные словари с их тщательным подбором материала остаются важным источником сведений о лексической сочетаемости. Статья посвящена разработке базы данных сочетаемости, в которую вошли словосочетания двух типов: статистические и лексические. Первая группа представлена коллокациями, которые были получены на основе корпусов текстов, в том числе при помощи статистических методов, в то время как лексические (или словарные) коллокации включают те, которые были зафиксированы в словарях русского языка и представляют собой так называемый золотой стандарт. В статье мы обратимся именно ко второму типу и проиллюстрируем принципы создаваемого ресурса на примере атрибутивных словосочетаний, построенных по модели «прилагательное + существительное».

**2. Обзор лексикографических проектов**

Для русского языка существует несколько лексикографических справочников (преимущественно в печатном виде), в которых представлена информация о сочетаемости. Прежде всего это толковые словари, например, БАС [Большой академический словарь русского языка 2004–], МАС [Словарь русского языка 1999] и БТС [Большой толковый словарь русского языка 1998], в которых устойчивые словосочетания выделены в отдельные части словарных статей или получили специальную разметку.

Еще одну группу лексикографических источников представляют собой специализированные словари. Уникальным проектом является Толково-комбинаторный словарь русского языка [Мельчук, Жолковский 1984], в котором устойчивая сочетаемость описана при помощи лексических функций. Словарь [Регинина, Тюрина, Широкова 1980] предназначен для иностранных учащихся, в нем представлено около 3 тыс. устойчивых словосочетаний. Словарь устойчивых глагольно-именных сочетаний русского языка [Дерибас 1983] также нацелен на изучающих русский язык и содержит более 5 тыс. словосочетаний, большинство из которых представлено биграммами. Словарь усилительных словосочетаний русского и английского языков [Убин 1987] имеет две части (прямую и реверсивную, т.е. словарные статьи представлены для главных и зависимых слов) и содержит более 10 тыс. единиц. Словарь коллокаций [Борисова 1995] 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 [Денисов, Морковкин 2002] представлены 2 500 словарных статей для существительных, глаголов и прилагательных. Авторами делается различие между лексической и семантической сочетаемостью, 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 [Апресян 2014-2017], который включает обширную информацию о сочетаемости, отраженную отдельно в словарных статьях.

Также доступен ряд проектов, которые нацелены на описание сочетаемости и представлены в виде базы данных или в виде онлайн систем. Если говорить об электронных ресурсах, то для русского языка НКРЯ [Национальный корпус русского языка] предоставляет ряд инструментов (n-грамм поиск со статистической оценкой, списки устойчивых слов и словосочетаний, лексические графы), а также на его основе были разработаны словари (например, [Кустова 2008]), в которых представлена ограниченная сочетаемость.

Среди остальных проектов для русского языка можно назвать FrameBank [Lyashevskaya 2010], который включает описание валентностных рамок для глаголов и конструкций. Ресурс “Collocations, Colligations, Constructions” [Kopotev et al. 2015] предоставляет информацию о сочетаемости на основе корпусов. В нем представлено два интерфейса в зависимости от подготовки пользователей. Еще одним проектом в данном направлении является семантический словарь Lexicograph [База данных “Lexicograph”].

**3. Описание базы данных**

В ходе исследования были рассмотрены следующие источники: 1) толковые словари (МАС и БТС); 2) словари сочетаемости [Борисова 1995; Мельчук, Жолковский 1984; Регинина, Тюрина, Широкова 1980; Убин 1987]; 3) электронный словарь [Кустова 2008]. На основе данных словарей было извлечено около 17 тыс. атрибутивных словосочетаний.

При анализе толковых словарей нами были извлечены данные из заромбовой части словарных статей, так как именно в ней представлена информация об ограниченной сочетаемости (о фразеологических оборотах и об устойчивых сочетаниях). В других словарях коллокации были представлены более простым образом: списком или отдельными словарными статьями.

C проектом можно ознакомиться на сайте «База данных коллокаций»[[2]](#footnote-2). Одноименная база данных была реализована с помощью платформы для создания онлайн-ресурсов, и в ней доступны следующие возможности:

* поиск коллокаций по главному (опорному) слову;
* поиск коллокаций по коллокату;
* визуализация;
* просмотр каждой отдельной коллокации и её лингвостатистических характеристик;
* просмотр ссылок на печатные и электронные издания словарей;
* просмотр ссылок на корпусы текстов.

Поскольку база данных может быть востребована разными группами пользователей, было решено разработать два интерфейса, которые подразумевают лингвистический и статистический поиск.

Первый тип интерфейса дает возможность просмотреть коллокации, выданные по запросу для главного слова или для коллоката. Результаты содержат список словосочетаний, в котором представлена следующая лингвистическая информация: определение лемм из Викисловаря, тип синтаксической структуры (например, прилагательное + существительное, глагол + существительное, наречие + прилагательное и пр.), ссылка на пример употребления в НКРЯ, наличие/отсутствие словосочетания в корпусах СинТагРус и Тайга, пересечение с другими словарными коллокациями. Нами также был введен словарный индекс — количество словарей, в которых зафиксирована коллокация. Соответственно, чем он больше (максимальное значение в случае атрибутивных словосочетаний равно 6), тем больше вероятность того, что словосочетание является воспроизводимым в речи и, как следствие, его необходимо запомнить (если речь идет о пользователе, изучающем русский язык). В табл. 1 представлено количество коллокаций, соответствующих каждому словарному индексу.

*Таблица 1.* **Словарный индекс**

|  |  |
| --- | --- |
| **Словарный индекс** | **Количество коллокаций** |
| 6 | 1 |
| 5 | 25 |
| 4 | 145 |
| 3 | 668 |
| 2 | 4803 |
| 1 | 11843 |

Таким образом при рассмотрении словосочетаний, которые зафиксированы в нескольких словарях, можно говорить о формировании частотного списка подобных коллокаций, который, однако, будет основан не на частоте, а на воспроизводимости в словарях и, соответственно, в речи. К этим коллокациям относятся, например, сочетания с прилагательными *глубокий* (*глубокая благодарность*, *глубокий интерес*, *глубокое удовлетворение*), *острый* (*острая борьба, острая дискуссия, острая полемика*) и *широкий* (*широкий выбор, широкая известность, широкое сотрудничество*). Поскольку словарные коллокации включают единицы, длина которых превышает два слова, то для них также дополнительно показано пересечение с другими коллокациями. Например, *бросить беглый взгляд* имеет пересечение с двумя словосочетаниями: *бросить взгляд* и *беглый взгляд*.

Для более интуитивно понятного отражения результатов предусмотрено графическое представление результатов запроса. Для визуализации коллокаций использовались библиотека NetworkX для Python 3.7. Ребра графов представлены в шести цветах, каждый цвет соответствует общему словарному индексу:

* желтый цвет — словарный индекс равен 1;
* синий цвет — словарный индекс равен 2;
* фиолетовый цвет — словарный индекс равен 3;
* красный цвет — словарный индекс равен 4;
* коричневый цвет — словарный индекс равен 5;
* черный цвет — словарный индекс равен 6.

На рис. 1 показан пример визуализации для лексемы «буйный». Таким образом, примерами словосочетаний, которые были найдены только в одном словаре, являются *буйный спор*, *буйная головушка*, *буйный океан*, *буйное цветение*. В то время как в двух словарях были зафиксированы следующие коллокации: *буйный натиск*, *буйный смех*, *буйная фантазия*, *буйное воображение*, *буйный рост* и др.

Каждый узел в визуализированной структуре соответствует лемме, при этом цвет узла обозначает определённую часть речи. Например, в рамках рассматриваемых нами атрибутивных словосочетаний синий цвет (на рисунке светло-серый, например, *восторг*, *океан*, *цветение*) обозначает имена существительные, а оранжевый — имена прилагательные (в данном случае только одно слово *буйный*).

С увеличением количества единиц базы данных встал вопрос о том, что с некоторыми частотными словами связано много коллокатов, поэтому такие примеры (например, в них могут встречаться по 100 и более слов) могут представлять сложность при восприятии пользователем. В качестве порогового значения было выбрано значение равное 35 примерам, т.к. стандартная визуализация NetworkX показывает для них удобочитаемые результаты. Поэтому для слов с числом коллокатов больше порогового было принято решение демонстрировать только так называемые значимые коллокации, под которыми мы понимаем те, которые обладают словарными индексами от 2 и выше, т.е. встречающиеся не менее чем в двух словарях.



*Рис. 1*. **Визуализация для лексемы «буйный»**

Статистический поиск подразумевает более специализированные возможности отображения результатов, которые нацелены на продвинутых пользователей.



*Рис. 2.* **Фрагмент выдачи результатов статистического интерфейса для лексемы «восторг»**



*Рис. 3.* **Фрагмент выдачи результатов статистического интерфейса для лексемы «восторг» (продолжение)**

Каждая словарная запись о коллокации снабжена следующей статистической информацией (см. рис. 2 и рис. 3):

* информация о вхождении той или иной коллокации в словари (всего 9 словарей);
* общий словарный индекс коллокации;
* относительная частота в ipm на основе НКРЯ и корпуса Araneum Russicum Maximum;
* значения мер ассоциации на основе Araneum Russicum Maximum (MI, MI3, log-likelihood, logDice, t-score).

Дополнительно для демонстрации значений метрик была использована инфографика (см. пример результатов выдачи для лексемы «аргумент» на рис. 4). Значения разных статистических коэффициентов показаны в виде столбцов диаграмм. На рис. 4 наибольшие значения меры t-score (равные 119 и 118) соответствуют словосочетаниям *весомый аргумент* и *главный аргумент*.



*Рис. 4.* **Инфографика для лексемы «аргумент»**

Различные статистические метрики могут указывать на неслучайный характер связи между компонентами словосочетаний.

**4. Заключение**

В заключение необходимо отметить, что база данных нуждается в пополнении, а также снабжении примерами из корпусов текстов не только в виде гиперссылок, но также в виде типовых контекстов. Планируется, что пользователи смогут отметить на их взгляд удачные коллокации или им будет доступна возможность добавить новые.

Разрабатываемая база данных может быть использована при обучении русскому как иностранному, для создания приложений, связанных с автоматической обработкой текстов, и при составлении словарей.

**Литература**

1. Апресян Ю.Д. (2014-2017), Активный словарь русского языка. Т. 1-3. Москва: «Языки славянской культуры».

2. База данных “Lexicograph”. URL: http://lexicograph.ruslang.ru (дата обращения: 23.05.2021).

3. Большой академический словарь русского языка в 30 т. (2004–), М.: Российская академия наук. Институт лингвистических исследований, Наука.

4. Большой толковый словарь русского языка: А-Я / Сост., гл. ред. канд. филол. наук С.А. Кузнецов (1998), СПб.: Норинт.

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

6. Денисов П.Н., Морковкин В.В. (2002), Словарь сочетаемости слов русского языка: 3-е изд., испр. М.

7. Дерибас В.М. (1983), Устойчивые глагольно-именные словосочетания русского языка. Москва: Русский язык.

8. Кустова Г.И. (2008), Словарь русской идиоматики. Сочетания слов со значением высокой степени. М.. URL: http://dict.ruslang.ru/magn.php (дата обращения: 23.05.2021).

9. Мельчук И.А., Жолковский А.К. (1984), Толково-комбинаторный словарь современного русского языка. Опыты семантико-синтаксического описания русской лексики. Вена: Wiener Slavistischer Almanach.

10. Национальный корпус русского языка. URL: http://ruscorpora.ru (дата обращения: 23.05.2021).

11. Регинина К.В., Тюрина Г.П., Широкова Л.И. (1980), Устойчивые словосочетания русского языка: Учеб. пос. для студентов-иностранцев / Под ред. Л. И. Широковой М.

12. Словарь русского языка: В 4-х т. / Под ред. А. П. Евгеньевой. 4-е изд., стер. (1999), М.: Рус. яз.

13. Убин И.И. (1987), Словарь усилительных словосочетаний русского и английского языков. М.: Рус. яз.

14. Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. (2015), 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, pp. 43-45.

15. 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.

**References**

1. Apresyan Yu.D. (2014-2017), Aktivnyy slovar’ russkogo yazyka. Vol. 1-3. Yazyki slavyanskoy kul’tury, Moscow.

2. Baza dannykh“Lexicograph”. URL: http://lexicograph.ruslang.ru (date of access: 23.05.2021).

3. Bol’shoy tolkovyy slovar’ russkogo yazyka (1998), S.A. Kuznetsov (ed.). Norint, St. Petersburg.

4. Bolshoy akademicheskiy slovar v 30 tomakh. (2004–2016). Nauka, Moscow, St. Petersburg.

5. Borisova E.G. (1995), Slovo v tekste. Slovar’ kollokatsiy (ustoychivykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov. Filologiya, Moscow.

6. Denisov P.N., Morkovkin V.V. (comp. and ed.-in-chief) (2002), Slovar’ sochetaemosti slov russkogo yazyka. Moscow: Russkij jazyk.

7. Deribas V.M. (1983), Verb-Noun Collocations in Russian. Moscow: Russian language. Ustojchivyje glagol’no-imennyje slovosochetanija russkogo jazyka. Moskva: Russkij jazyk.

8. Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. (2015), 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, pp. 43-45.

9. Kustova G.I. (2008), Slovar’ russkoj idiomatiki. Sochetanija slov so znachenijem vysokoj stepeni, available at: http://dict.ruslang.ru/magn.php (date of access: 23.05.2021).

10. 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.

11. Mel’chuk I., Zholkovsky A. (1984), Tolkovo-kombinatornyy slovar russkogo yazyka. Vienna.

12. Natsional’ny korpus russkogo yazyka. URL: http://ruscorpora.ru (date of access: 23.05.2021).

13. Oubine I. (1987), Slovar’ usilitel’nykh slovoso-chetaniy russkogo I angliyskogo yazykov. Russian Language, Moscow.

14. Reginina K.V., Tjurina G.P., Shirokova L.I. (1980), Ustoychivye slovosochetaniya russkogo yazyka: Uchebnoye posobiye dlya studentov-inostrantsev. Shirokova L.I. (ed.). Moscow.

15. Slovar’ russkogo yazyka v 4 tomakh (1999), Yevgen’yeva A. P. (ed.-in-chief). Vol. 1–4, 4th edition, revised and supplemented. Russkiy yazyk, Moscow.

**BUILDING RUSSIAN COLLOCATIONS DATABASE:
A CASE STUDY OF GOLD STANDARD
FOR ADJ-NOUN COLLOCATIONS**

M. Khokhlova, I. Mamaev

**Abstract.** The study of collocability is an important task and is still highly relevant in linguistics. The present paper deals building the Russian Collocations Database which includes both verified or dictionary collocations, as well as data from text corpora extracted automatically. The authors describe adj-noun collocations that can be found in the database and in what number of dictionaries they were registered. Also the paper discusses methods that were used for collocation visualization and presents the results of their evaluation against the collected data from verified resources.

**Keywords**. Collocations, database, dictionaries, statistical methods, Russian language

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Хохлова Мария Владимировна**

Санкт-Петербургский государственный университет (Россия).

**Khokhlova Maria**

St. Petersburg State University (Russia).

***E-mail: m.khokhlova@spbu.ru***

**Мамаев Иван Дмитриевич**

Санкт-Петербургский государственный университет (Россия).

**Mamaev Ivan**

St. Petersburg State University (Russia).

***E-mail: st079541@student.spbu.ru***

1. «Исследование выполнено за счет гранта Российского научного фонда (проект № 19-78-00091)». [↑](#footnote-ref-1)
2. https://collocations.spbu.ru/ [↑](#footnote-ref-2)