О САМОМ СЛОЖНОМ: ИЗУЧЕНИЕ СОЧЕТАЕМОСТИ СЛОВ ОНЛАЙН

М.В. КОПОТЕВ
mkopotev@hse.ru
д-р философии,
профессор Национального исследовательского университета «Высшая школа экономики»,
лектор Хельсинкского университета
Санкт-Петербург, Россия
Хельсинки, Финляндия

Ключевые слова:
лексическая и грамматическая сочетаемость, коллокации, коллигации, коллострукции, корпус

DOI: 10.37632/PI.2020.283.6.005

Статья посвящена описанию онлайн-ресурса CoCoCo (cococo.cosyco.ru), который позволяет получать информацию о лексической и грамматической сочетаемости слов. Педагогическая цель справочника состоит в том, чтобы дать быстрые ответы на самые сложные вопросы при изучении РКИ: как строить речь не только по правилам, но и идиоматично. Для этого используются сложные статистические расчеты, основанные на большом материале трех корпусов русского языка.

Часто студенты, имеющие большой объем знаний о грамматике и лексике изучаемого языка, не могут использовать его идиоматически, как это делают носители языка. Правильное сочетание слов с трудом поддается обобщению, на него часто не хватает времени на занятиях, а словари на дают ответы на все возникающие вопросы. Все это делает преподавание идиоматики на занятиях по русскому языку непростой задачей. На помощь приходят современная лингвистическая теория и компьютерные технологии.

Первым шагом в системном изучении идиом, или фразеологизмов, является разработанная в начале XX в. классификация фразеологических единиц[1]. Основы этой классификации были заложены в работах швейцарских лингвистов, в частности Шарля Балли (1909, рус. перевод [2]). В русской лингвистике пионером в новой области стал И.Е. Аничков, опубликовавший в 1920-е гг. серию статей, позже ставших основой его диссертации. Именно он ввел в научный обиход термин *идиома* («по аналогии с терминами *фонема, морфема, синтагма, семема*» [1: 108]) и обосновал роль идиоматики как науки. В отличие от подхода, получившего в дальнейшем широкое распространение в русистике, И.Е. Аничков не противопоставлял свободные и несвободные сочетания: «Словосочетания разной степени прочности, компоненты которых в разной мере являются переменными, не исключая так называемых «свободных сочетаний», – одинаково реальны и объективны <...>. Рассмотрению с этой точки зрения подлежит весь язык, а не только, как представлял Ш. Балли, часть языка, названная им «фразеологией» [1: 285]. В отечественном языкознании бульшую известность получили работы другого последователя Ш. Балли – В.В. Виноградова. В его работах середины 1940-х гг. [6, 7] была адаптирована классификация швейцарского ученого и заложены основы известного разделения фразеологизмов на сращения, единства и сочетания. Эта классификация дожила до наших дней, включив в себя еще одну группу единиц – фразеологические выражения (см. [Шанский 1985]). Важным шагом в исследовании идиоматики стали работы И.А. Мельчука [14, 15], в которых он представил более строгое определение устойчивости и предложил общую классификацию, выведя ее за пределы собственно лексических единиц – фразеологизмов. Эта классификация опирается на совмещение русской и американской лингвистических традиций (прежде всего указанных работ Виноградова и [21, 27, 32]). Позднее исследователь предложил типологию *фразем*, существующих

[1] Подробный обзор современных подходов можно найти в монографии [3: 9–24].

на трех уровнях языковой системы – лексемном, морфологическом и синтаксическом. Отталкиваясь от принятого определения идиомы, согласно которому «семантика идиомы не складывается из семантики входящих в нее элементов» (некомпозициональность семантики), исследователь определил фразему как словосочетание, означаемое и означающее которого не могут быть построены без ограничений и регулярно. При множестве различных подходов и теорий, существующих в рамках современной лингвистики, общим является признание языковой способности частным случаем когнитивного навыка, который возникает из ситуации использования языка. Этот подход полностью согласуется и с переориентацией методики преподавания языка на коммуникативные подходы.

Среди всех типов фразеологических единиц меньше всего внимания было уделено тем, которые наиболее часто встречаются в нашей речи: семантически прозрачным, но устойчивым выражениям. Такие полуоформленные единицы И.А. Мельчук назвал *полуфраземами*, в англоязычной литературе они получили название статистических *коллокаций* (от англ. *collocations*). В современных работах в рамках грамматики конструкций и корпусной лингвистики к этим единицам добавились *коллигации* (англ. *colligations*) и *коллострукции* (англ. *collostuctions*). Все три типа единиц в определенной степени пересекаются, и все обладают общим свойством: совместной встречаемостью конституирующих элементов. Под полуфраземами, или **статистическими коллокациями**, обычно подразумевают «неслучайное сочетание двух и более лексических единиц, характерное как для языка в целом (текстов любого типа), так и определенного типа текстов (или даже (под)выборки текстов)» [20: 575]. Примерами коллокаций могут служить сочетания «крепкий чай» или «греет душу»[2]. Под **коллигациями** в научной литературе понимают неслучайное сочетание грамматических и лексических параметров, или «совокупности морфолого-синтаксических условий, обеспечивающих сочетаемость языковых единиц» [9: 36; 30: 15]. Примерами коллигаций могут служить как простые случаи согласования существительного и прилагательного, так и более сложные случаи, например выбор вида глагола после модального слова. **Коллострукциями** обычно называют сочетания, в которых связанными оказываются и семантические, и грамматические параметры [12, 31], например: существительные singularia tantum со значением ‘ягоды’ после глаголов обладания *собирать, покупать,* например *собирать клубнику, малину*.

Для выявления таких единиц в тексте корпусная лингвистика использует специальные статистические инструменты, которые основываются на предположении, что частота сочетания должна быть более значимой, чем у каждой из входящих в нее единиц по отдельности. Для измерения совместной встречаемости используются специальные статистические инструменты (см. [5, 11, 18, 25, 29]). В последние десятилетия наблюдается всплеск интереса к этой теме и в преподавании языков, связанный с возросшей ролью корпусной лингвистики [8, 10, 16, 17, 26]. В то же время использование корпусов в учебном процессе часто оказывается трудной задачей в силу сложности интерфейса, неясности инструкций, а часто и большого количества корпусного материала, который невозможно проработать на занятии.

Справочник сочетаемости слов CoCoCo (cococo.cosyco.ru)

Цель проекта CoCoCo – создание нового технического средства обучения, которое решает как раз эту задачу: изучение сочетаемости слов с учетом лексических и грамматических параметров [22–24]. Основной результат проекта – онлайн-сервис, который поощряет студентов к активному «исследованию» и к участию в процессе обучения в отличие от более традиционного освоения языка через накопление пассивного знания. В этом контексте тип обучения, к которому мы стремимся, можно назвать интерактивным. Ресурс был разработан исследовательской группой на базе Хельсинкского университета под руководством автора этой статьи и рассчитан специально на преподавателей и студентов РКИ и учитывает их потребности, навыки работы и даже возможное недостаточное знание языка. Справочник позволяет с помощью многоступенчатого статистического анализа выявлять сочетаемостные предпочтения лексем, располагая признаки на единой шкале от более к менее устойчивым. Основные цели создания справочника:

– обеспечение легкого и быстрого «входа» как для преподавателя, так и для студента с неполным знанием языка;

[2] Это расширенное понимание коллокаций несколько противоречит более строгому, собственно лингвистическому, пониманию коллокаций как единиц, имеющих связанное, некомпозициональное значение (ср. [4, 13, 28] и др.). С другой стороны, такой подход позволяет включить широкий и, надо сказать, слабо оформленный список единиц, предполагающий дальнейшую более строгую классификацию, исходящую не из теоретических предпосылок, а из закономерностей, выявляемых в реальном массиве языковых данных.

– использование больших корпусных данных, позволяющее осуществлять поиск как в ограниченной тематической области, так во всем корпусе;

– анализ сочетаний слов с использованием статистических методов (англ. *corpus-driven approach*);

– создание удобной системы обучения, генерирующей ответы на индивидуальные запросы пользователя.

Лежащая в основе технология позволяет определить, какие параметры в цепочке слов, грамматические и/или лексические, наиболее тесно связаны. Например, после глагола *читать* мы ожидаем существительное, из всех грамматических признаков которого только падеж зависит от глагола. Это простой случай синтаксического управления, или коллигации. Другой пример – слово *баклуши*, которое, естественно, не контролирует ни один грамматический параметр глагола, однако лексически тесно связано с конкретным глаголом *бить*. Это пример лексической зависимости, или коллокации. В более сложных случаях связанными оказывается не отдельная лексема или грамматический параметр, а целый класс лексем с определенной грамматической и лексической семантикой. В таком случае созданная система предлагает список таких словоформ, например: а) *играть в футбол, баскетбол, хоккей*, б) *играть в игрушки, куклы, машинки* (Ср. англ. *to play soccer, basketball, hokey vs to play with toys, with dolls, with cars*).

Структура справочника

К сожалению, не существует большого и аккуратно аннотированного корпуса для русского языка. В справочнике CoCoCo используется три корпуса, имеющие свои достоинства и недостатки. Во-первых, это подкорпус со снятой омонимией НКРЯ, который предлагает качественное морфологическое аннотирование, однако его объем довольно мал – всего 5 млн слов. Два более объемных корпуса – I-RU и «Тайга» – содержат больше языковых данных, около 140 и 400 млн слов соответственно, однако качество автоматического морфологического аннотирования в этих корпусах ниже. Корпус «Тайга» дополнительно разделен на 5 жанровых подкорпусов, 80 млн слов в каждом (поэзия, социальные сети, субтитры, журналы и новости), что позволяет исследовать жанрово обусловленную сочетаемость слов.

Тексты из указанных корпусов были разделены на так называемые *н-граммы*, или фрагменты длиной от одного до пяти слов. Это, с одной стороны, позволяет производить необходимые вычисления, а с другой – эффективно хранить большой объем информации. Как результат, интерфейс справочника позволяет искать как двухсловные сочетания, так и цепочки слов длиной до пяти слов. Искомое слово может располагаться не только справа от заданной цепочки, но и слева и посередине, например: какие прилагательные стабильно появляются в сочетании *«в ADJ конце»*? Ответ: *самом, дальнем, противоположном, другом, любом.*

Поиск может быть осуществлен в двух режимах: по словоформе и по начальной форме (лемме) с уточнением морфологических параметров. Например, чтобы узнать, с какими существительными чаще всего используется словосочетание *искать в*, нужно указать, что поиск ведется по леммам (т.е. будут учтены все формы глагола *искать* и предлога *в*); в этом случае в поле ввода нужно задавать начальные формы (Рис. 1). В третьем поле необходимо ввести грамматические параметры: в нашем примере В.п. существительного. Для выбора морфологических параметров необходимо задать часть речи и нажать кнопку справа от части речи.

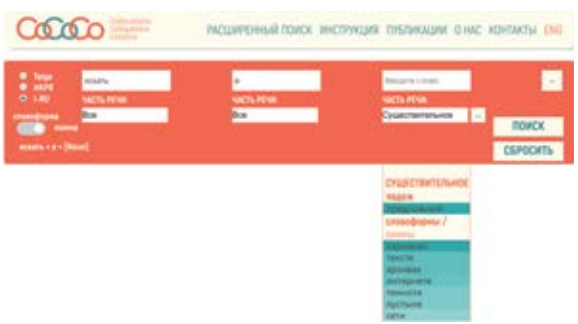

Рис. 1

Морфологическая разметка каждого корпуса имеет свои особенности, например, в корпусе I-RU выделяются всего шесть базовых падежей, тогда как в двух других корпусах отдельно отмечены формы второго родительного, звательного и др. Детальную информацию об особенностях разметки можно найти на сайтах разработчиков корпусов:

– «Тайга» tatianashavrina.github.io/taiga_site,

– НКРЯ: ruscorpora.ru/en/corpora-morph.html,

– I-RU: smlc12.leeds.ac.uk/itweb/help/Russian_Language_Tutorial.pdf.

В результатах поиска отражаются не все слова, встретившиеся в текстах корпуса, но только самые значимые в заданном контексте – именно на них и надо обратить внимание при изучении языка. Результаты выводятся в виде таблицы, в которых указываются самые устойчивые

словоформы, сочетающиеся с наречием *рано* (Рис. 2), или лексемы/леммы (Рис. 3).

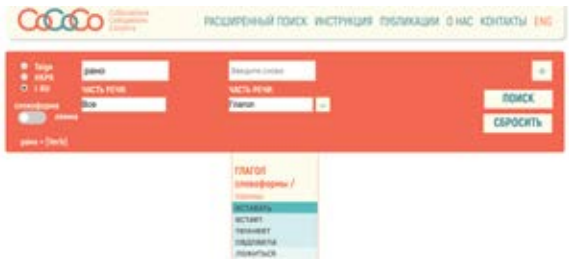

Рис. 2. Устойчивые глагольные словоформы, сочетающиеся с наречием «рано»

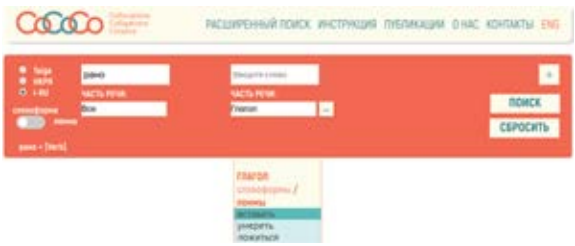

Рис. 3. Устойчивые глагольные леммы, сочетающиеся с наречием «рано»

Для того чтобы выяснить коллигационные (т.е. морфологические) предпочтения, достаточно задать слово в одном или нескольких полях и выполнить поиск. Выведенная таблица с результатами покажет самые важные грамматические параметры для каждой части речи. Например, в сочетании «*до* + существительное» сам предлог управляет двумя падежами – родительным и вторым родительным (Рис. 4).

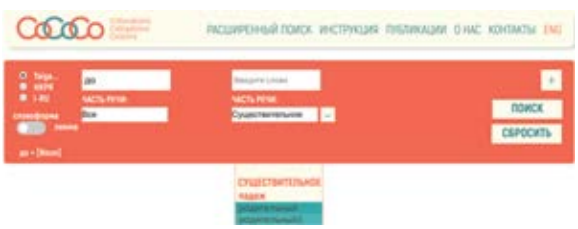

Рис. 4. Падежное управление предлога «до»

Однако сочетание «*не до* + существительное» показывает, что наиболее ожидаемым падежом существительного в этом сочетании является второй родительный, поскольку часто встречается в сочетаниях *не до жиру, не до смеху* (Рис. 5). Как мы видим, пользователь получает информацию не просто о падежном управлении предлога, но и о более специфической грамматической сочетаемости.

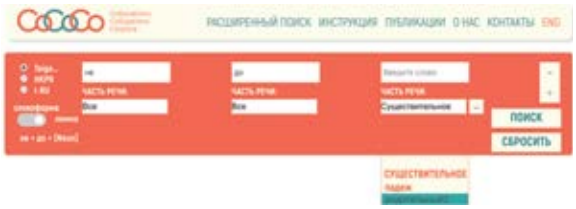

Рис. 5. Падежное управление сочетания «не до»

Наконец, последний тип сочетаемостных возможностей, коллострукции, позволяет понять, с какими семантическими классами типично сочетается слово или группа слов. Покажем это на примере двух подкорпусов корпуса «Тайги». Нажав на знак многоточия справа от названия корпуса «Тайга», пользователь увидит список подкорпусов. Далее зададим поиск «Прилагательное в И.п. + существительное *снег* в И.п.» в поэтическом корпусе, как на Рис. 6:

Рис. 6. Запрос «Прилагательные + существительное «снег» в поэтическом подкорпусе»

Помимо коллигаций и коллокаций последняя таблица с результатами поиска выведет прилагательные, наиболее характерные для этого типа текстов, сгруппированные в пять семантических классов (Рис. 7).

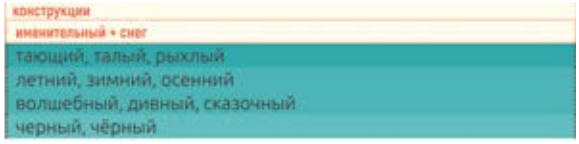

Рис. 7. Устойчивые прилагательные при существительном «снег» в поэтическом подкорпусе «Тайги»

Если произвести такой же поиск в новостном подкорпусе, алгоритм подсчитает, что с существительным *снег* сочетаются прилагательные всего одной и семантически совсем другой группы (Рис. 8).

Рис. 8. Устойчивые прилагательные при существительном «снег» в новостном подкорпусе «Тайги»

Следует еще раз подчеркнуть, что система показывает не просто самые частые комбинации лексических и грамматических параметров, но определяет те, которые являются самыми устойчивыми, типичными в заданном запросе.

Практическое применение

Справочник CoCoCo в первую очередь необходим для быстрого поиска ответов на многочисленные вопросы, которые, с одной стороны, не описаны в достаточной степени в словарях, а с другой – требуют серьезных усилий и знаний при обращении к обычному интерфейсу любого корпуса. Однако ресурс может быть использован и на занятиях или в качестве основы для домашних заданий. Ниже приведены примеры заданий, которые можно выполнять на основе справочника.

Задание 1. С какими прилагательными сочетаются следующие существительные:

живот _____

лоб _____

глаз _____

голова _____

Комментарий. В любом языке слова, обозначающие части тела, образуют лингвоспецифичные сочетания, в том числе метафорические. Выполнение этого задания позволит не только узнать идиоматическую сочетаемость, например *впалый живот*, но и понять, какие метафоры связаны с определенной частью речи, а заодно уточнить, что слово *голова* в русском языке может быть и существительным мужского рода в значении ‘руководитель’: *городской, сельский голова*.

Задание 2. Какими предлогами управляют следующие глаголы:

отозваться _____

отчитываться _____

предостеречь _____

Комментарий. Это задание даст возможность выяснить сочетаемость относительно редких глаголов, информация о которых часто отсутствует или не полно представлена в существующих словарях.

Задание 3. С какими глаголами сочетаются следующие наречия:

сурово _____

косо _____

вручную _____

врасплох _____

Комментарий. Из всех частей речи наречия и наречные формы являются одним из самых идиоматически связанных с контекстом. Об этом говорит и словообразовательная история многих из них, и небольшие, часто закрытые, списки глаголов, к которым они способны примыкать.

Задание 4. С какими семантическими классами сочетаются следующие леммы:

ходить по _____

бродить по _____

путешествовать по _____

колесить по _____

Комментарий. Слова, близкие по значению, часто отличаются друг от друга оттенками смысла, которые трудно уловить студенту-иностранцу, а также разными сочетаемостными возможностями, которые понять и выучить гораздо легче. Задания такого рода позволяют сформировать устойчивые знания о сочетаемости и, возможно, понять семантические различия между сочетаниями.

Описанный в статье ресурс является прежде всего прикладным, образовательным проектом. Возможности его применения ограничены только фантазией пользователя и тем материалом, который представлен на сайте. В то же время этот ресурс позволяет поставить и более теоретический вопрос, одинаково актуальный и для лингвистов, и для преподавателей. Соссюровское противопоставление *langue* и *parole* постепенно заменяется представлением о первичности речевой деятельности, строго иерархичная языковая система меняется на вероятностную шкалу устойчивости от речевого штампа до грамматического правила. При таком понимании языка его адекватным описанием оказывается не грамматика и словарь, взятые отдельно, а компьютерная база данных, в которой хранятся не правила, а тренды. ■

ЛИТЕРАТУРА

1. Аничков И.Е. Труды по языкознанию. СПб., 1997.
2. Балли Ш. Французская стилистика. М., 1961.
3. Баранов А.Н., Добровольский Д.О. Аспекты теории фразеологии. М., 2008.
4. Борисова Е.Г. Коллокации. Что это такое и как их изучать. М., 1995.
5. Браславский П.И., Соколов Е.А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог», 2006.
6. Виноградов В.В. Основные понятия русской фразеологии как лингвистической дисциплины // Труды юбилейной сессии. Серия филологических наук. Л., 1946.
7. Виноградов В.В. Русский язык: (Грамматическое учение о слове). М.; Л., 1947.
8. Влавацкая М.В. Система базовых понятий комбинаторной лингвистики // МНКО. 2011. № 1. cyberleninka.ru/article/n/sistema-bazovyh-ponyatiy-kombinatornoy-lingvistiki-1.
9. Влавацкая М.В. Комбинаторная лингводидактика в контексте современного языкового образования // Сибирский педагогический журнал. 2015. № 4.
10. Клочихин В.В. Формирование коллокационной компетенции обучающихся на основе электронного лингвистического корпуса // Вестник ТГУ. 2019. № 179.
11. Кочеткова Н.А. Статистические языковые методы. Коллокации и коллигации // Новые информационные технологии в автоматизированных системах. 2013. № 16. cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii.
12. Кузнецова Ю.Л., Велейшикова Т.В. Современные корпусные исследования языка: новые подходы // Вопросы языкознания. 2010. № 6.
13. Кустова Г.И. Словарь русской идиоматики. Сочетания слов со значением высокой степени. 2008. dict.ruslang. ru/magn. php.
14. Мельчук И.А. О терминах «устойчивость» и «идиоматичность» // Вопросы языкознания. 1960. Т. 4.
15. Мельчук И.А. Об одном классе фразеологических сочетаний // Проблемы устойчивости фразеологических единиц. Тула, 1968.
16. Павельева Т.Ю. Изучение коллокаций на основе лингвистических корпусов текстов // Вестник ТГУ. 2016. № 3–4.
17. Петросян И.В. Обучение коллокациям современного английского языка // Вестник СПбГУ. Серия: Язык и литература. 2014. № 2.
18. Хохлова М.В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia. 2008. Т. 34.
19. Шанский Н.М. Фразеология современного русского языка. М., 1985.
20. 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. Т. 2.
21. Jackendoff R. The boundaries of the lexicon // Idioms: Structural and psychological perspectives. 1995. Vol. 133.
22. Kopotev M., Pivovarova L., Kochetkova N., Yangarber R. Automatic detection of stable grammatical features in n-grams // Papers from the 9th Workshop on Multiword Expressions (MWE 2013). Workshop at NAACL 2013 (Atlanta, Georgia, USA), June 13–14, 2013. Atlanta, 2013.
23. Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. CoCoCo: Online Extraction of Russian Multiword Expressions // The 5th Workshop on Balto-Slavic Natural Language Processing, 10–11 September 2015, Hissar, Bulgaria. Sofia, 2015.
24. Kopotev M., Pivovarova L., Kormacheva D. Constructional generalization over Russian collocations // Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching. Mémoires de la Société Néophilologique de Helsinki / B. Sanromán Vilas. Helsinki, 2016.
25. Kormacheva D., Pivovarova L., Kopotev M. Evaluation of collocation extraction methods for the Russian language // Quantitative approaches to the Russian language. Routledge, 2018.
26. Lackman K. Teaching Collocations: Activities for Vocabulary Building. Toronto, 2011.
27. Makkai A. Idiom structure in English. Hague. 1972.
28. Mel'cuk I. Phrasemes in language and phraseology in linguistics // Idioms: Structural and psychological perspectives. 1995.
29. Pecina P. An extensive empirical study of collocation extraction methods // Proceedings of the ACL Student Research Workshop. 2005.
30. Sinclair J. Corpus, concordance, collocation. Oxford, 1991.
31. Stefanowitsch A., Gries S. Th. Collostructions: investigating the interaction between words and constructions // International journal of corpus linguistics. 2003. 8 (2).

32. Weinreich U. Problems in the analysis of idioms // Substance and structure of language. 1969. Vol. 23.

References

1. Anichkov I.E. Trudy po yazykoznaniyu. SPb., 1997.
2. Balli Sh. Francuzskaya stilistika. M., 1961.
3. Baranov A.N., Dobrovol'skij D.O. Aspekty teorii frazeologii. M., 2008.
4. Borisova E.G. Kollokacii. Chto eto takoe i kak ih izuchat'. M., 1995.
5. Braslavskij P.I., Sokolov E.A. Sravnenie chetyrekh metodov avtomaticheskogo izvlecheniya dvuhslovnyh terminov iz teksta // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Trudy Mezhdunarodnoj konferencii «Dialog», 2006.
6. Vinogradov V.V. Osnovnye ponyatiya russkoj frazeologii kak lingvisticheskoj discipliny // Trudy yubilejnoj sessii. Seriya Filologicheskih nauk. L., 1946.
7. Vinogradov V.V. Russkij yazyk: (Grammaticheskoe uchenie o slove). M.; L., 1947.
8. Vlavackaya M.V. Sistema bazovyh ponyatij kombinatornoj lingvistiki // MNKO. 2011. № 1. cyberleninka.ru/article/n/sistema-bazovyh-ponyatiy-kombinatornoy-lingvistiki-1.
9. Vlavackaya M.V. Kombinatornaya lingvodidaktika v kontekste sovremennogo yazykovogo obrazovaniya // Sibirskij pedagogicheskij zhurnal. 2015. № 4.
10. Klochihin V.V. Formirovanie kollokacionnoj kompetencii obuchayushchihsya na osnove elektronnogo lingvisticheskogo korpusa // Vestnik TGU. 2019. № 179.
11. Kochetkova N.A. Statisticheskie yazykovye metody. Kollokacii i kolligacii // Novye informacionnye tekhnologii v avtomatizirovannyh sistemah. 2013. № 16. cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii.
12. Kuznecova Yu.L., Velejshikova T.V. Sovremennye korpusnye issledovaniya yazyka: novye podhody // Voprosy yazykoznaniya. 2010. № 6.
13. Kustova G.I. Slovar' russkoj idiomatiki. Sochetaniya slov so znacheniem vysokoj stepeni. 2008. dict. ruslang.ru/magn. php.
14. Mel'chuk I.A. O terminah «ustojchivost'» i «idiomatichnost'» // Voprosy yazykoznaniya. 1960. T. 4.
15. Mel'chuk I.A. Ob odnom klasse frazeologicheskih sochetanij // Problemy ustojchivosti frazeologicheskih edinic. Tula, 1968.
16. Pavel'eva T.Yu. Izuchenie kollokacij na osnove lingvisticheskih korpusov tekstov // Vestnik TGU. 2016. № 3–4.
17. Petrosyan I.V. Obuchenie kollokaciyam sovremennogo anglijskogo yazyka // Vestnik SPbGU. Seriya: Yazyk i literatura. 2014. № 2.
18. Hohlova M.V. Eksperimental'naya proverka metodov vydeleniya kollokacij // Slavica Helsingiensia. 2008. T. 34.
19. Shanskij N.M. Frazeologiya sovremennogo russkogo yazyka. M., 1985.
20. Yagunova E.V., Pivovarova L.M. Priroda kollokacij v russkom yazyke. Opyt avtomaticheskogo izvlecheniya i klassifikacii na materiale novostnyh tekstov // Nauchno-tekhnicheskaya informaciya. Seriya 2: Informacionnye processy i sistemy. 2010. T. 2.
21. Jackendoff R. The boundaries of the lexicon // Idioms: Structural and psychological perspectives. 1995. Vol. 133.
22. Kopotev M., Pivovarova L., Kochetkova N., Yangarber R. Automatic detection of stable grammatical features in n-grams // Papers from the 9th Workshop on Multiword Expressions (MWE 2013). Workshop at NAACL 2013 (Atlanta, Georgia, USA), June 13–14, 2013. Atlanta, 2013.
23. Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. CoCoCo: Online Extraction of Russian Multiword Expressions // The 5th Workshop on Balto-Slavic Natural Language Processing, 10–11 September 2015, Hissar, Bulgaria. Sofia, 2015.
24. Kopotev M., Pivovarova L., Kormacheva D. Constructional generalization over Russian collocations // Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching. Mémoires de la Société Néophilologique de Helsinki / B. Sanromán Vilas. Helsinki, 2016.
25. Kormacheva, D., Pivovarova L., Kopotev M. Evaluation of collocation extraction methods for the Russian language // Quantitative approaches to the Russian language. Routledge, 2018.
26. Lackman K. Teaching Collocations: Activities for Vocabulary Building. Toronto, 2011.
27. Makkai A. Idiom structure in English. Hague. 1972.
28. Mel'cuk I. Phrasemes in language and phraseology in linguistics // Idioms: Structural and psychological perspectives. 1995.
29. Pecina P. An extensive empirical study of collocation extraction methods // Proceedings of the ACL Student Research Workshop. 2005.

30. Sinclair J. Corpus, concordance, collocation. Oxford, 1991.
31. Stefanowitsch, A., Gries, S.Th. Collostructions: investigating the interaction between words and constructions // International journal of corpus linguistics. 2003. 8 (2).
32. Weinreich U. Problems in the analysis of idioms // Substance and structure of language. 1969. Vol. 23.

M.V. Kopotev
Higher School of Economics – National Research University
University of Helsinki
Saint Petersburg, Russia
Helsinki Finland

THE MOST DIFFICULT ONE: TEACHING WORD CO-OCCURRENCES ONLINE

Lexical and grammatical co-occurrence, collocations, colligations, collostructions, corpus.

The article is devoted to the description of the online resource CoCoCo (cococo.cosyco.ru), which allows a user to receive information about both lexical and grammatical co-occurrences of words. The goal of the resource is to provide quick answers to the most difficult questions in learning Russian: how to produce speech not just by rules, but also idiomatic. To achieve this, sophisticated statistical calculations are applied, based on a large language data extracted from three Russian corpora.

НОВОСТИ НОВОСТИ НОВОСТИ НОВОСТИ

Британский словарь английского языка Collins English Dictionary объявил главным словом 2020 г. «локдаун» – режим изоляции, введенный в связи с пандемией коронавируса. Российские коллеги из Государственного института русского языка им. А.С. Пушкина назвали главными словами 2020 г. в России наш аналог «локдауна» – «самоизоляцию» и слово – зеркало исторических событий в стране – «обнуление». Ведущие филологи России предложили варианты слов года на Всероссийской онлайн-конференции, посвященной актуализации, расширению и обновлению списка словарей, грамматик и справочников, содержащих нормы современного русского литературного языка при его использовании в качестве государственного языка Российской Федерации.

Конференция проходила в формате видеосвязи, поэтому в личный список участников – российских ученых, авторов учебников и словарей, педагогов – попали слова «дистанцирование», «удаленка», «дистант», «цифровизация». Д-р филол. наук, замдиректора по научной работе Института лингвистических исследований РАН Марина Приемышева предложила в «фавориты года» название интернет-платформы Zoom в российском варианте написания – «зум». По ее словам, в Институте лингвистических исследований РАН лидерами в списке слов года стали «ковид» и «коронавирус», у каждого из них за время распространения нового вируса появилось до 150 наименований в словообразовательных гнездах.

Термин «пандемия» объявили словами 2020 г. член Совета по русскому языку при Президенте Российской Федерации и Правительственной комиссии по русскому языку Константин Деревянко, его коллега по Правкомиссии и директор Института русского языка им. В.В. Виноградова РАН Мария Каленчук и профессор кафедры русского языка СПбГУ Сергей Кузнецов. «Мое слово года – непечатное, но оно присутствует у всех русскоязычных людей в сознании», – пошутил автор программы и учебников по русскому языку для начальной школы Станислав Иванов, но после согласился отдать личное первое место «ковиду». В Институте Пушкина в финал также вышли слова, которые активно использовались в 2020 г.: «голосование», «конституция», «поправки».

По материалам сайта gramota.ru