

---

# КроссЛексика: Готовая версия

---

**Большаков Игорь Алексеевич**

Независимый исследователь, Москва

*[iabolshakov@gmail.com](mailto:iabolshakov@gmail.com)*

**Гельбух Александр Феликсович**

Профессор национального политехнического института, Мехико

*[gelbukh@gelbukh.com](mailto:gelbukh@gelbukh.com)*

# Общая характеристика (1/3)

- **КроссЛексика – самое большое в мире собрание всех типов связей между русскими словами.** Сугубо компьютерных словарей, сопоставимых с КЛ по структуре, размеру словника и количеству связей между элементами словника, не существует ни для одного иного языка в мире, даже английского.
- КЛ предназначена для самого **широкого круга пользователей**, включая ученых (в том числе – лингвистов), преподавателей, инженеров, журналистов, бизнесменов, дипломатов, военных, учащихся, студентов, аспирантов, пенсионеров, домохозяек.
- Лингвистическая база КроссЛексики основана на десятках словарей, сотнях других печатных источников и многих тысячах запросов к интернет-поисковикам Гугл и Яндекс.

# Общая характеристика (2/3)

- КЛ поддерживается в **двух основных версиях**:
  - ❑ Сокращенная по объему словника и количеству связей интернет-версия с ограниченным пользовательским интерфейсом, дающая доступ с компьютера под любой операционной системой.
  - ❑ Интерактивная полнообъемная версия, автономно воплощаемая на компьютерах под ОС Windows, с полным набором современных интерактивных средств.
- **Обе версии КЛ позволяют**:
  - ❑ обучаться русскому языку во всех сферах его использования и с полным учетом лексики, морфологии и синтаксиса,
  - ❑ совершенствовать русские тексты (документы, статьи, брошюры, книги) в процессе их создания за компьютером.

# Общая характеристика (3/3)

## Полная интерактивная версия:

- Содержит в несколько раз больше слов и межсловных связей и позволяет создавать разнообразные производные словари, как чисто лингвистические (синонимов, паронимов, моделей управления и др.), так и тематические (по кулинарии, медицине, обработке документации и др.)
- Поддерживает английский интерфейс, содержит двусторонний англо-русский словарь для ввода запросов по-английски и нахождения переводов русских слов и оборотов.
- Обеспечивает удобный доступ к мировому информационному фонду через три важных поисковика.

На июнь 2016 г. предлагается в свободном доступе

**сокращенная интернет-версия** КроссЛексики:

<http://www.xl.gelbukh.com>

---

# Сопоставительные ресурсы для русского языка

## ➤ в печатной форме:

- ❑ **Тихонов** А.Н., Тихонова Е.Н., Тихонов С. А., Чупашева О. М., Зуева М. Ю. Комплексный словарь русского языка. Москва, 2007.
- ❑ **Морковкин** В. В., Богачева Г. Ф., Луцкая Н. М. Большой универсальный словарь русского языка. Москва, 2016.

## ➤ в интернете:

- ❑ Словарь ассоциаций Reright: [reright.ru](http://reright.ru)
  - ❑ Национальный корпус русского языка: [ruscorpora.ru](http://ruscorpora.ru)
-

---

# Сопоставительные ресурсы для других языков

Для **английского** языка

в печатной и электронной форме:

- Oxford Collocation dictionary for students of English
  - Oxford University Press, 2009
  - Online Oxford Collocation dictionary
- Macmillan Collocations Dictionary for Learners of English
  - Macmillan Publishers Ltd, 2016
  - [www.macmillandictionary.com](http://www.macmillandictionary.com)

Для **испанского** языка

DiCE: Diccionario de Colocaciones del Español

[www.dicesp.com/paginas](http://www.dicesp.com/paginas)

Для **французского** языка

González Rodríguez, Toni. Dictionnaire des collocations

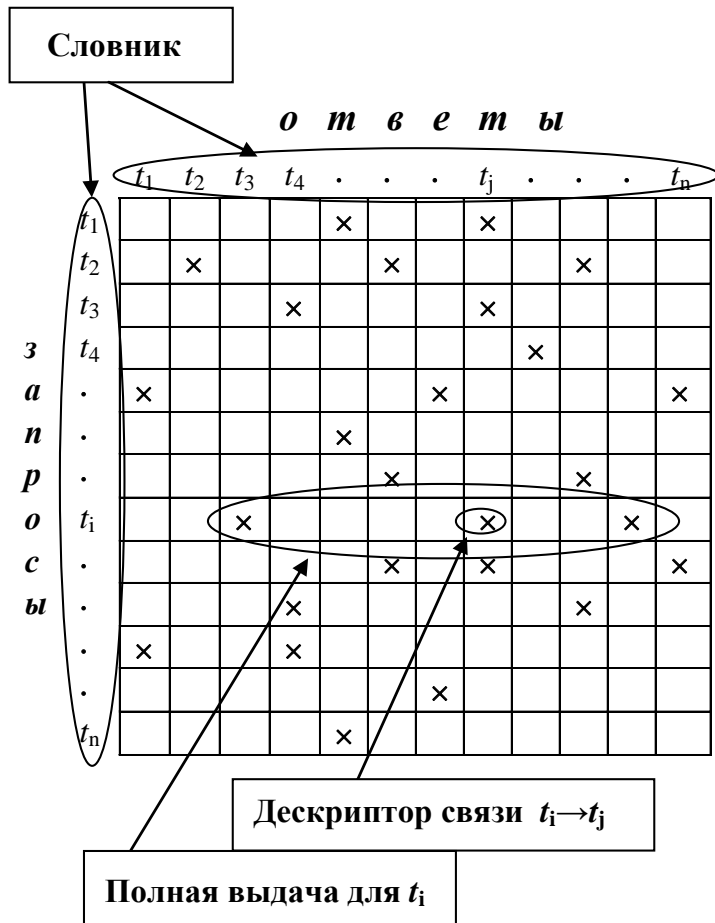
[www.tonitraduction.net](http://www.tonitraduction.net)

---

# КроссЛексика – единственный сугубо компьютерный словарь

- Если объект ИС состоит из нескольких частей, то он должен быть доступен через любую его часть в одинаковом виде. Так, коллокация *наблюдать за ссорой* должна быть найдена по обоим коллокатам: *наблюдать* и *ссора*.
- Если объект ИС имеет варианты с разными связями, то эти варианты должны быть размечены (пронумерованы) по всей БД . У объектов-респондентов должно быть указано, какой конкретно вариант имеется в виду. Так, если в словаре не разметить омонимы, то будут признаны правильными коллокации типа *симпатичное доверенное лицо*.
- Коррекция объекта в ИС должна производиться синхронно, где бы этот объект в БД ни находился. Выполнить это требование трудно, если просто оцифровать печатный словарь страница за страницей.
- Компьютерный словарь должен предоставлять свои ресурсы в реальном времени как человеку-пользователю, так и программам, связанным с этим словарем.

# Глобальная структура КроссЛексик: Гигантская матрица {Словник x Словник}



- Элемент словника = вокабула
- Элементы матрицы – дескрипторы связи между запрошенной  $t_i$  и ответной вокабулой  $t_j$   
 $i, j = 1, \dots, 310000+$
- Связи ограничены языком и реалиями внешнего мира. Из 100 миллиардов ячеек матрицы непушта лишь каждая 10000-я.
- У вокабулы в среднем 32 связи.



# Вокабулы относятся к четырем частям речи

## ➤ **Субстантивные:**

- Отдельное существительное (42%): *абжур, битва, бифштекс, благо, блины...*
- Именное словосочетание: *алкогольные напитки, ближнее зарубежье, сельское хозяйство, точка зрения, уровень жизни, экономический рост...*

## ➤ **Глагольные** в инфинитиве или личных формах:

- Одиночный глагол (54%): *говорить, идти, обсуждать, спать, ругать...*
- Глагольный оборот: *навести страх, оказывать внимание, испытать ужас...*

## ➤ **Адъективные:**

- Отдельное прилагательное: *абстрактный, авансовый, автономный, авантюрный...*
- Отдельное причастие: *зادвинутый, мытый, перевезенный, желающий...*
- Согласованный адъективный оборот: *бросающийся в глаза, хорошо одетый...*
- Несогласованный адъективный оборот: *бойцовской породы, большой дальности, в елочку...*
- Наречие или ограничительная частица в синтаксической роли определения существительного: *только, исключительно, фактически...*

## ➤ **Адвербиальные:**

- Отдельное наречие: *абсолютно, абстрактно, адски, аляповато, быстро...*
- Отдельное деепричастие: *базируясь, надев, торопясь, шепча...*
- Адвербиальный оборот: *аккуратным образом, более или менее, как выжатый лимон, в особой степени, куда попало, мелкой дрожью...*

# Первая цель КроссЛексики (1/2)

- покрытие максимально широкой аудитории

- **Политематичность**, т.е. как можно более полное покрытие областей использования языка: экономики, финансов и бизнеса; социальной и политической сферы; гуманитарных наук и религии; точных и естественных наук; инженерии и технологий; строительства и архитектуры; медицины; экологии; спорта; кулинарии; бытового языка включая бранную лексику.
- **Концентрация информации** из академических и специальных словарей, из потока новостей, политической, экономической и научной аналитики в интернете, из объявлений, гламурных журналов по знаменитостям, моде, туризму, автомобилям.
- Включение как **лингвистической**, так и **энциклопедической** информации.

# Первая цель КроссЛексики (2/2)

## - покрытие максимально широкой аудитории

- Упрощенная, но **наглядная система помет** при словах и коллокациях. Например, особенности **стиля** размечаются как *нейтральное, специальное/книжное, разговорное, вульгарное и неграмотное*. Тем самым КЛ затрагивается как бытующая, так и нормативная стороны языка. КЛ предостерегает пользователя от «некультурных» выражений типа *более лучше* или *болтнуть глупость*.
- **Толерантный интерфейс:**
  - ❑ Включение графических опций новых слов, например, *бренд и брэнд, плеер и плейер*.
  - ❑ Необязательность знания чисто лингвистических понятий и терминов. Например, для секции выдачи можно пользоваться названием *Одноклассники* вместо *Когипонимы*.
- **Упрощенческие отклонения от лингвистического канона.** Например, два разных вида глагола и два числа существительного рассматриваются как разные слова; причастия и деепричастия отделены от глагола и отнесены к прилагательным и наречиям соответственно.

# Вторая цель КроссЛексики

– встраивание пользователя в современный информационный мир

- Предоставление **английского интерфейса**, что учитывает современную лингвистическую ситуацию. Встроенный **двусторонний англо-русский словарь** позволяет автоматически переводить запросы с английского языка, получать переводы русских слов и коллокаций на английский и многих английских коллокаций на русский.
- Конструирование запросов к **трем важнейшим поисковикам** интернета путем выбора из словника или имеющегося набора коллокаций – для получения более подробной информации по выбранному слову, обороту или проблеме.

# Принципы построения КроссЛексики

(1+2+3/5)

- **Включение трех известных типов связей** между вокабулами:
  - ❑ **Семантико-синтагматических** (далее синтаксических)
  - ❑ **Семантико-парадигматических** (далее семантических)
  - ❑ **Паронимических** (т.е. буквенного или морфного сходства).
- **Покрытие всех типов синтаксических связей** между четырьмя главными частями речи – существительными, глаголами, прилагательными и наречиями. Все типы связей рассматриваются в обе стороны, так что в выдаче предусмотрены 22 разноименных секции, никогда не покрываемые полностью.
- **Сетевой принцип**: вокабула включается в словарь вместе со всеми обнаруженными на текущий момент ее связями. (Иные словари предпочитают линейный принцип: простую последовательность статей словника.)

# Принципы построения КроссЛексики

(4+5/5)

- **Декомпозиционный принцип:** любое однозначное слово или словосочетание, входящее в многословную вокабулу, само является отдельной вокабулой, и можно проследить путь декомпозиции, например,  
*авиа- и железнодорожный транспорт = авиационный транспорт + железнодорожный транспорт*  
*авиационный транспорт = авиационный + транспорт*  
*железнодорожный транспорт = железнодорожный + транспорт*
- **Языковая многоуровневость:** кроме указанных связей синтаксического и семантического характера для каждой вокабулы дается ее **морфологическая парадигма**.  
Существительные склоняются по шести стандартным и двум дополнительным падежам; глаголы представлены инфинитивом и 12 личными формами; для прилагательных учитываются все значения рода, числа, падежа и одушевленности.

# Коллокации (= словосочетания)

- Коллокация – это совокупность двух полнозначных вокабул, синтаксически связанных в тексте и устойчиво совместимых по смыслу.
- Коллокации бывают частотными и редкими, свободными и фразеологическими, включают разнообразные предикативные выражения.
- В синтаксической связи между двумя полнозначными вокабулами может стоять **служебное слово**, предлог или сочинительный союз *и / или / да*:  
*полнозн. слово1 → (служебное слово) → полнозн. слово2*  
*сотрудничество → ради → мира*
- Каждая коллокация доступна **с двух сторон**.

# Типы многочисленных коллокаций (сотни тысяч) (1/2)

- Определительная пара **существительное – прилагательное**:  
*краснокочанная капуста, явный наглец, полная ясность...*
- **Глагол – его прямое / косвенное / предложное дополнение-существительное**  
(включая ходовые обстоятельства):  
*рассмотреть вопрос, ковырять в носу, остаться из-за погоды, купить на рынке, отличаться сдержанностью...*
- **Причастие / прилагательное – его прямое / косвенное / предложное дополнение-существительное**  
(включая ходовые обстоятельства):  
*рассмотревший вопрос, ковырявший в носу, оставшийся из-за погоды, красный от гнева, купленный на рынке, отличающийся нравом,...*
- Определительная пара **глагол / прилагательное / наречие – наречие**:  
*резко высказаться, полностью ясный, ужасно страшно...*



# Типы многочисленных коллокаций (сотни тысяч) (2/2)

- **Подлежащее-существительное – сказуемое в виде личной формы глагола или краткого адъектива:**

*самолет вылетел, внимание (было / будет) привлечено, доклад (был / будет) краток, враг напал, глазки бегают, категоричность смущает, детсад закрылся ...*

- **Существительное – подчиненное ему существительное:**

*сердце матери, наложение взыскания, отличия в произношении, борьба против терроризма...*

- **Деепричастие / наречие – его прямое / косвенное / предложное дополнение-существительное:**

*рассмотрев вопрос, ковыряя в носу, оставшись из-за погоды, купив на рынке, отличаясь сдержанностью, близко от города,...*

# Некоторые типы малочисленных словосочетаний (единицы или десятки тысяч)

- **Устойчивые сочиненные пары:**  
*автобусы и троллейбусы, ясный и четкий, экономический и культурный, быть или не быть, взвесить и решить, власть и бизнес, в срок и в полном объеме, базы и склады, наука и техника, авиа- и железнодорожный транспорт, десять заповедей и семь смертных грехов...*
- **Глагол – его инфинитивное дополнение:**  
*собраться поехать, мечтать выкупаться, хотеть перекусить...*
- **Существительное – его инфинитивное дополнение:**  
*соблазн сказать, желание уйти, проблема выжить...*
- **Прилагательное / причастие – его инфинитивное дополнение:**  
*готовый действовать, желающий начать, агитирующий голосовать...*
- **Наречие / деепричастие – его инфинитивное дополнение:**  
*абсурдно полагать, безумно ввязываться, берясь изучать, боясь возразить, бессмысленно идти...*

# Семантические связи

## Самые многочисленные:

### ➤ **Синонимы**

### ➤ **Семантические дериваты**

Простой пример группы СД:

{ *извлечение; извлекать, извлечь; извлеченный, извлекающий, извлекавший; извлекая, по извлечению, путем извлечения* }

↑ **Здесь встречаются элементы канонических морфопарадигм** ↑

↑ **и дается основное множество энциклопедических сведений** ↑

## Менее многочисленные:

➤ **Когипонимы.** Пример: *мясо* → *вырезка, грудинка, гуляш, котлеты, фарш...*

➤ **Ассоциации.** Пример: *аденоиды* → *аллергия, бассейн, гланды, гомеопатия, кашель, лазеротерапия, миндалины, слух...*

➤ **Меронимы / холонимы**

➤ **Гипонимы / гиперонимы**

➤ **Антонимы**

Все указанные связи хорошо известны, кроме ассоциаций. Они извлекаются из сочиненных пар в запросах к Рунету и в БД поисковиков.

# Вокабулы с наибольшим числом ассоциаций в Рунете

558	<i>беременность</i>	125	<i>человек</i>
264	<i>здоровье</i>	122	<i>любовь</i>
257	<i>алкоголь</i>	121	<i>бизнес</i>
172	<i>спорт</i>	121	<i>курение</i>
143	<i>диабет</i>	120	<i>дети</i>
136	<i>диета</i>	117	<i>культура<sub>1</sub></i>
131	<i>цены</i>	112	<i>похудение</i>
127	<i>мужчины</i>	104	<i>религия</i>

# Использование семантических связей

- СемСы помогают понять **СМЫСЛ** вокабул.  
Примеры:  
**Synonym** (*граффити*) = *настенная живопись*  
**Synonym** (*графт*) = *трансплантат*  
**Synonym** (*халяльный*) =  
*отвечающий мусульманским нормам*  
**Hyperonym** (*эндометриоз*) =  
*акушерско-гинекологическая болезнь*
- СемСы помогают построить словосочетания, в КЛ отсутствующие. Пример:  
(**Hyperonym**(*каллы*) = *цветы*) & (*букет цветов*)  
→ (*букет калл*)
- СемСы отражают многочисленные **энциклопедические сведения**

---

# Энциклопедические сведения

- Названия геообъектов: континентов, океанов, морей, горных цепей...
  - Названия крупнейших городов мира в привязке к странам
  - Сведения о 60 иностранных государствах (по 20 ведущим – более подробные)
  - Названия и другие сведения о десятках городов и регионов России
  - Около 300 наиболее частых русских имен вместе с их диминутивами
  - Имена ряда известных политических, деловых, научных и культурных деятелей мира
  - Названия ряда крупных организаций и корпораций мира
  - Названия ряда известных художественных произведений мира
  - Терминология точных, естественных и гуманитарных наук, спорта, кулинарии; медицинская терминология
-

# Основное приложение – диалоговое

- **Предпосылка:** Пассивное знание языка у многих заметно шире активно используемых языковых средств. Если показать, как можно выразить ту же мысль иначе, пользователь легко найдет более подходящий и идиоматичный вариант.
- **Характер работы:** пользователь вводит запрос и использует выдачу
  - ❑ для углубленного изучения русских слов и выражений или
  - ❑ для редактирования создаваемого текста с целью придать ему правильную и идиоматичную форму

# Примеры лингвистических справок

- Как можно выразиться глаголом о плате за *проезд*?  
– *оплатить / оплачивать проезд* либо *платить / заплатить за проезд* (*проплатить проезд* и *оплатить за проезд* тоже включены, но снабжены пометами разговорности ● и запрета ● соответственно)
  - Как «запустить» иск? – Можно *внести / возбудить / вчинить / подать / предъявить иск*, а также *обратиться с иском*.
  - Как еще можно назвать *бразильских женщин*? – *бразильянки*. А как *иракских женщин*? – Да никак иначе! (Но *иракец, иракцы* среди семантических дериватов есть)
-



---

# Различение паронимов

## ■ вероятный

является определением для:

адрес

альтернатива

вариант

версия

визит

встреча

выбор

гипотеза

запасы

изменение

.....

## ■ вероятностный

является определением для:

автомат

алгоритм

анализ

анализатор

аспекты

вывод

задача

идеи

контроль

логика

.....

---

# Одноименными зонами выдачи формируются лингвистические словари

- Морфологический словарь
- Словарь определительных словосочетаний
- Словарь глагольно-именных словосочетаний
- Словарь моделей управления для глаголов
- Словарь моделей управления для существительных
- Словарь синонимов
- Словарь морфемных паронимов
- Словарь буквенных паронимов
- Словарь смысловых ассоциаций
- Словарь семантических дериватов
- и другие

# Пример: Словарь моделей управления для глаголов

- Содержит списки «валентностей» 14 тыс. глаголов, иллюстрированных 460 тыс. примеров.
- Количество беспредложных и предложных «валентностей» у глагола иногда непривычно:
  - 28 – *организовать*
  - 23 – *установить*
  - 13 – *изучать*
  - 12 – *проверять*
  - 11 – *соорудить*
  - 9 – *обеспечивать*
  - 6 – *анализировать*

---

# Для любого специального подсловника можно «вырезать» тематический словарь

- **КроссЛексика-чиновник** – для широкого круга работников сферы управления.
  - **КроссЛексика-гурман** – для любителей поесть или приготовить что-то самому.
  - **КроссЛексика-пациент** – для широкого круга обычных людей, часто посещающих врачей.
  - **КроссЛексика-красотка** – для девушек и дам, желающих видеть себя здоровыми и красивыми.
-

# Недиалоговые приложения (не входят в словарь)

Внешняя программа обращается к словарю через специальную утилиту КЛ и использует выдачу самостоятельно. Примеры:

- Автоматическое обнаружение и исправление смысловых ошибок типа *истерический центр* или *неутомимый голод* – с помощью коллокаций и паронимов КЛ.
- Разрешение неоднозначности омонимов по контексту – с помощью коллокаций КЛ.
- Лексическая фильтрация результатов синтаксического разбора – с помощью коллокаций КЛ.
- Стеганография и стеганализ – с помощью коллокаций и синонимов КЛ.

# Глобальные параметры двух версий на май 2016

- **Вокабул** **310 тыс. (138 тыс.)**
  - ❑ Существительных 46%
  - ❑ Глаголов 13%
  - ❑ Прилагательных 24%
  - ❑ Наречий и др. ч. р. 16%
- **Связей между** **9.98 млн. (6.33 млн.)**
  - ❑ Синтаксических **5.68 млн. (4.28 млн.)**
    - из них:
      - определительных 29%
      - глагольно-именных 24%
  - ❑ Семантических **3.32 млн. (1.72 млн.)**
  - ❑ Паронимических **0.97 млн. (0.33 млн.)**

# Несколько сравнений (1/2)

- Большой универсальный словарь русского языка (В. В. Морковкин и др., 2016):
  - ❑ Около 450 тыс. русских словосочетаний
  - ❑ 30 тыс. словарных статей
- Oxford Collocation Dictionary for students of English (Oxford, 2009):
  - ❑ 250 тыс. английских словосочетаний
  - ❑ 9 тыс. словарных статей
- КроссЛексика, полная интерактивная версия (2016):
  - ❑ 2.84 млн. русских словосочетаний
  - ❑ 130 тыс. вокабул в словосочетаниях

# Несколько сравнений

(2/2)

СМ – словарь Морковкина    СА – словарь ассоциаций Reright

Существительное                    СМ                    СА                    КЛ

сравнение по числу связей

<i>беременность</i>	0	193	1709
<i>здоровье</i>	152	0	1087
<i>алкоголь</i>	0	206	557
<i>спорт</i>	93	192	523
<i>диета</i>	0	210	419
<i>аборт</i>	0	196	148

сравнение по числу определений

<i>человек</i>	653	71	1394
<i>лицо1</i>	164	64	978
<i>женщина</i>	28	64	888
<i>глаза</i>	59	64	842
<i>политика</i>	74	64	635
<i>шутка</i>	77	64	192



---

**Вопросы направляйте по адресу**  
**[iabolshakov@gmail.com](mailto:iabolshakov@gmail.com)**

**и посмотрите сокращенную**  
**интернет-версию**

**<http://www.x1.gelbukh.com>**

---