КроссЛексика — большой электронный словарь сочетаний и смысловых связей русских слов

Crosslexica: a large electronic dictionary of collocations and semantic links between russian words

Большаков И. А. (bolshakov34@mail.ru)

Национальный политехнический институт, Мехико, Мексика

Большой русский электронный словарь содержит словник из 185 тыс. титулов, 1,75 млн. словосочетаний, 2 млн. смысловых связей между словами, английские переводы титулов, их морфопарадигмы. Работает в диалоге (редактирование текстов, обучение языку) и доступен из программ парсинга, разрешения омонимии, обнаружения/исправления смысловых ошибок, стеганографии.

1. Введение

За последние 20 лет в русском письменном языке произошли существенные сдвиги.

- Изменилась и пополнилась лексика. Накапливавшиеся ранее разговорные слова и жаргонизмы выплеснулись на страницы изданий, на телевидение, в Интернет. Появилось много новых заимствований.
- Соответственно изменился и пополнился состав словосочетаний, которыми, по формулировке И. Мельчука [6], только и говорит человек.
- В части владения языком ситуация поляризовалась. На одном полюсе, возросло число авторов (обозревателей, журналистов, ученых-гуманитариев), виртуозно владеющих языком и не стесненных советскими речевыми штампами. На другом полюсе, появилась масса полуграмотных «афторов», демонстрирующих Интернете убогую приблатненную или англизированную лексику и попирающих нормативную орфографию. Между этими полюсами сохранилась группа научно-технических авторов, не блещущих стилем и разнообразием лексики, но обеспечивающих языковую преемственность в своей сфере.

В итоге академические словари русского языка заметно устарели. Изданные в последние годы «большие» словари, напр., [1, 2, 3, 5], успевают истолковывать новации, но не отображают увеличенной массы словосочетаний.

В те же десятилетия радикально усовершенствовалась вычислительная техника. Типовой объем дискового накопителя увеличился в тысячи раз. В памяти десктопа, лаптопа, мобильника уже умещаются словари любого объема. При выдаче словарных статей на экран уже не обязательно повторять их привычный бумажный формат.

Данный доклад отражает результаты 19-летней работы над электронным русским словарем КроссЛексика, который объединяет подъязыки разных групп пользователей без навязывания норм и лексических крайностей. Словарь

- содержит около 185 тысяч слов и неразрывных выражений, с особым упором на их 1,75 млн. сочетаний;
- отражает 2 млн. смысловых связей (синомимы, антонимы, гипонимы-гиперонимы, меронимы-холонимы, семантические дериваты);
- включает более 0,5 млн. паронимических связей (буквенного либо морфемного сходства);
- тематически универсален, т.е. содержит политическую, научную, экономическую, политехническую и общежитейскую лексику;
- комбинаторикой различает слова из нескольких тысяч омонимических групп;
- с помощью отношений синонимии и род-вид оперативно порождает более 1,5 млн. словосочетаний, в словаре непосредственно не представленных;
- управляет порядком и полнотой выдачи и по требованию отсеивает ненужную данному пользователю информацию;
- дает английские переводы для титулов словника и частотных словосочетаний;
- приводит морфопарадигмы большинства титулов словника.

Структура КроссЛексики состоит из алфавитно упорядоченного словника и матрицы классифици-

рованных связей между его элементами. Грамматически правильное сочетание восстанавливается через связь синтагматического типа, если ввести любой знаменательный компонент сочетания.

Основной режим работы КроссЛексики — диалоговый, в нем можно редактировать русские тексты, обучаться русскому языку, получать всевозможные лексические и грамматические справки. Можно обращаться к словарю и из внешних программ, осуществляющих парсинг, разрешающих омонимию, обнаруживающих и исправляющих смысловые ошибки и др.

Конкретно, в КроссЛексике представлена следующая тематика:

- Экономика и бизнес;
- Политика и политология;
- Различные разделы техники: радиоэлектроника, компьютеры, программирование, автомобили, бытовая техника, строительство и др.;
- Точные и естественные науки: математика, физика, химия, биология, география и др.;
- Гуманитарные науки (напр., лингвистика), искусство, религия;
- Медицина (преимущественно бытовая);
- Бытовой язык, включая бранную лексику без мата.

2. Общие параметры и лингвистическая информация словаря

Титулы словника делятся на существительные (31%), глаголы (21%), прилагательные (27%) и наречия (21%). Омонимических групп 2,3 тыс. с 5,4 тыс. разных смыслов. Раскрываются 3,6 тыс. склеек типа *физфак = физический факультет,* рассматриваемые и как титулы словника, и как словосочетания.

Суммарное количество словосочетаний — около 1,75 млн. (В [4] и [2] их примерно 100 тыс. и 200 тыс.) Количество семантических связей более 2 млн., паронимических связей — более 0,5 млн.

Титулы делятся на:

- Субстантивные (раздельно единств. и множеств. число);
- Глагольные (инфинитив + личные формы, два вида берутся раздельно);
- Адъективные (прилагательные или причастия двух видов раздельно);
- Адвербиальные (наречия или деепричастия двух видов раздельно).

Отказ от чисто лексемного принципа представления существительных и глаголов диктовался необходимостью отразить различия в комбинаторике подпарадигм. Титулы подпарадигм рассматриваются как взаимные семантические дериваты.

Служебные слова (предлоги, союзы) встроены в словосочетания и своих статей обычно не имеют. Предикативные высказывания типа *а пошел ты* отнесены к наречиям.

Субстантивная статья имеет титулом либо отдельное существительное (*абажур, абберация, аббревиатура, абзац, битва…*), либо устойчивое именное словосочетание (*алкогольные напитки, ближнее зарубежье, сельское хозяйство, точка зрения, уровень жизни, болеутоляющие средства…*).

Глагольная статья имеет титулом либо одиночный глагол (*говорить, идти, обсуждать, спать, демонстрировать…*), либо глагол с возвратным местоимением (*вести себя…*), либо глагольный оборот (*наводить страх, свалиться как подкошенный, залить фары, испытывать стремление…*).

Адъективная статья имеет титулом либо отдельное прилагательное (*абстрактный, авансовый, авантюрный, автономный, воздушно-реактивный…*), либо отдельное причастие, м.б. переходящее в прилагательное (*агонизирующий, вдвинутый, коррумпированный…),* либо адъективный оборот (*хорошо одетый, большой дальности, бросающийся в глаза, в елочку, как сталь, в денежном выражении, без подкладки, большого ума…*).

Адвербиальная статья имеет титулом либо отдельное наречие *(абстрактно, адски, долго, плохо, по-мужски, удовлетворительно…*), либо отдельное деепричастие *(базируясь, дрожа, надев, успев…),* либо адвербиальный оборот *(аккуратным образом, без воодушевления, более или менее, будто обухом по голове, как выжатый лимон, в особой степени, куда попало, мелкой дрожью, на цыпочках, долгое время…).*

Связи между статьями словника делятся на:

- **Синтагматические**, формирующие словосочетания;
- **Семантические**, связывающие слова со смысловым сходством;
- **Паронимические,** связывающие внешне сходные слова.

Словосочетание — это два знаменательные слова, синтаксически связанные и устойчиво совместимые по смыслу. В синтаксической связи между двумя знаменательными словами может стоять служебное слово (предлог или союз) согласно формуле

знамен. слово1 → (служебное слово) → → знамен. слово2,

например, *сотрудничество → ради → мира, пойти → на → курсы, уверенный → в → победе.*

Каждое словосочетание доступно с двух сторон. Доступ с одной из сторон дает одностороннюю связь, и таких связей в словосочетаниях ровно вдвое больше, чем самих словосочетаний.

Наиболее частотны в словаре следующие типы словосочетаний:

- **Существительное — прилагательное** или **глагол/прилагательное/наречие — наречие**, образующие определительные пары (*краснокочанная капуста, резко высказаться, полностью ясный, ужасно страшно*);
- **Причастие/прилагательное** — его прямое, косвенное или предложное **дополнение-существительное**, включая ходовые обстоятельства (*рассмотревший вопрос, ковырявший в носу, оставшийся из-за погоды, красный от гнева, купленный на рынке*);
- **Глагол** — его прямое, косвенное или предложное **дополнение-существительное,** включая ходовые обстоятельства (*рассмотреть вопрос, остаться из-за погоды, купить на рынке, отличаться сдержанностью*);
- **Деепричастие/наречие** — его прямое, косвенное или предложное **дополнение-существительное** (*рассмотрев вопрос, ковыряя в носу, купив на рынке, близко от города*);
- **Существительное-подлежащее —** его **сказуемое** в виде личной формы глагола или краткого прилагательного/причастия (*внимание (было) привлечено, доклад (был) краток, враг напал, глазки бегают*);
- **Существительное** — подчиненное ему **существительное** (*сердце матери, наложение взыскания, отличия в произношении, борьба с бюрократизмом*).

Менее частотны в словаре следующие типы словосочетаний:

- **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*);
- **Деепричастие/наречие** — его **инфинитивное дополнение** (*мечтая сказать, пожелав уйти, собравшись купаться*);
- **Глагол** — его **адъективное дополнение** (*вернуться здоровым, найти мертвым, считать выполненным*);
- **Прилагательное/причастие** — его **адъективное дополнение**: *найденный нормальным, вернувшийся здоровым*);
- **Деепричастие/наречие** — его **адъективное дополнение**: *найдя нормальным, считая выполненным*).
- Устойчивые **сочиненные пары** из одинаковых частей речи (*ясный и четкий, быть или не быть, власть и бизнес, в срок и в полном объеме, базы и склады, наука и техника*).

Статистика по базе словосочетаний выявляет следующие существительные с наибольшим числом управляющих глаголов:

514 *работа*1	387 *место*1	365 *руки*
463 *деньги*	377 *дом*1	339 *дело*1
413 *ребенок*	367 *дети*	333 *дорога*

Существительные с наибольшим числом определений:

1189 *человек*	548 *глаза*	520 *вид*1
715 *лицо*1	536 *женщина*	507 *режим*2
557 *работа1*	534 *взгляд1*	499 *голос*1

Глаголы с наибольшим числом дополнений:

2284 *быть*	1270 *стать*	963 *считать*
2185 *иметь*	1095 *начать*	959 *вести*
1442 *находиться*	1068 *получить*	951 *оказаться*

Прилагательные — наиболее частые определения:

2958 *большой*	1604 *новый*	1321 *явный*
2047 *крупный*	1463 *постоянный*	1202 *огромный*
1749 *небольшой*	1407 *полный*1	1183 *сильный*

Смысловые связи делятся на следующие группы:

- **Синонимы (**например, *дурак — болван*) в виде 19 тыс. синонимических групп в среднем по 5,6 элементов; односторонних синонимических связей 1,12 млн.
- **Семантические дериваты** — это группы типа {*Москва1, москвичи; московский…*} или {*извлечение; извлекать, извлечь; извлеченный, извлекший; извлекая, по извлечении, путем извлечения*}, односторонних связей 0,88 млн.
- **Часть/целое** (например, *террариум — зоопарк*), связей 21 тыс.
- **Род/вид** (например, *диплом1 — документ*), связей 14 тыс.
- **Антонимы** (например, *длинный — короткий*), связей 12 тыс. Они дополняются выдачей антонимов для синонимов данного слова и синонимов для антонимов.

Смысловые связи важны не только сами по себе. Во-первых, толковательные синонимы (а их в словаре много) способны разъяснить смысл слова, и это особенно важно при наличии глоссов только для омонимов. Так, у *кошерный* есть синоним *отвечающий иудейским нормам*. Во-вторых, с помощью семантически связанных слов можно построить из титулов словника новые правдоподобные словосочетания, непосредственно в базе не представленные, напр., получить новое словосочетание по формуле

(букет цветов) *&* ***(астры*** *IS_A* ***цветы)*** *→*
→ ***(букет астр)****.*

Паронимические связи в словаре представлены:

- **Буквенными** паронимами, т.е. словами той же части речи, отличающимися от данного слова на одну букву, например, *кадка*: {*кака, ка**с**ка, ка**ч**ка, ка**ш**ка, к**л**адка…*}.
- **Морфемными** паронимами, имеющими ту же часть речи и общий корень, но иное сочетание аффиксов, напр., {***бег, бег**ун, **бег**а, **бег**лость, при**беж**ище, про**беж**ка…*}.

Практически для каждого изменяемого титула словаря, включая многословные, дается его морфологическая парадигма.

У титулов словника есть английские переводы. Собранные вместе, они образуют отдельный словник, через который можно войти в русскую часть словаря.

Для словосочетаний введены три градации фигуральности (идиоматичности). Если пометы фигуральности нет, словосочетание понимается как есть: *идти в школу, вызвать слесаря*. При помете **idiom** словосочетание понимается только фигурально (идиоматически): *сесть в галошу, висеть на волоске.* При помете **mb idiom** словосочетание понимается либо фигурально, либо в прямом смысле: *сесть в лужу, первая ракетка*.

Большее число градаций имеет стиль (степень разговорности) слов и словосочетаний.

- Если пометы стиля нет, слово (словосочетание) является нейтральным и достаточно употребительным, так что его полезно знать (*стена, окно, книга, налоги…*).
- Помета в виде зеленого буллита указывает специальное, книжное или забытое слово (словосочетание). Предлагается пользоваться им, если нет опасности непонимания слушателями: *абсцесс, парадигма, экзистенциальный,…*
- Желтый буллит указывается чисто разговорное слово или выражение, которым предлагается не пользоваться в официальных документах: *мотать нервы, жевать сопли,….*
- Красный буллит указывается бранное слово или выражение, которым нельзя пользоваться при дамах, детях и в официальной обстановке: *говно, жопа, засранец, мудак, взять за яйца,…*
- Черный буллит указывает нейтральное бытующее выражение, смысл которого лингвисты рекомендуют передавать более нормативно: *оплатить за проезд, проплатить операцию.…*

Пользователю КроссЛексики предлагаются некоторые опции, а именно

- Можно выбрать язык обращения со словарем:
  - **Русский** (компоненты меню, названия разделов выдачи, толкования омонимов и справочная информация даются по-русски), либо
  - **Английский** (все вышеуказанное дается по-английски).
- В процессе работы с КроссЛексикой можно:
  - Выбрать алфавитный порядок выдачи основных типов словосочетаний либо частотный порядок (словосочетания с более частотными в словаре элементами выдаются первыми);
  - Установить порог отсечения словосочетаний с низкочастотными в словаре титулами;
  - Отменить выдачу бранной, разговорной и/или специальной лексики вместе с соответствующими словосочетаниями;
  - Ввести запрос (1) с клавиатуры, (2) подведя указатель к нужной строке в окне словника, (3) выбрав строку в обновляющемся списке *История,* либо (4) выбрав строку в списке словосочетаний на экране. Последний вариант начинает навигацию по словнику.

3. Приложения словаря

Приложения словаря возможны:

- **Диалоговые** (интерактивные), когда пользователь обращается к словарю в диалоговом режиме и использует результаты, например, при параллельном редактировании текста или при обучении русскому языку;
- **Недиалоговые** (неинтерактивные), когда внешняя программа обращается к словарю за информацией и использует результаты для своих целей.

Вот примеры диалоговых запросов русскоязычного пользователя:

- Как можно выразиться глаголом о *плате за проезд? платить, оплатить, оплачивать проезд* либо *заплатить за проезд* (*проплатить проезд* и *оплатить за проезд* тоже даются на экране, но снабжены черным буллитом).
- Как можно еще назвать *бразильских женщин?* — *бразильянки. А* как *иракских женщин*? — Да никак иначе! (Но *иракец, иракцы* допустимы.)
- Как «запустить» иск? — *внести, возбудить, вчинить, подать* или *предъявить иск,* а также *обратиться с иском.*
- Как управляет существительными глагол *забыть?*
  - **забыть что/кого?** *забыть адрес, багаж, вкус, времена, время, вчерашнее,…* (101 словосочетание)
  - **забыть о чем/о ком?** *забыть о времени, обо всем, о вчерашнем, о главном,…* (37)
  - **забыть про что/про кого?** *забыть про все, про главное, про детей, про диссертацию, про семью,…* (22)
  - **забыть в чем/в ком/где?** *забыть в вагоне, в гостях, в комнате, в кафе, в ресторане, в спешке…* (10),
  - **забыть на чем/на ком/где?** *забыть на диване, на кресле, на кровати…* (7)

  - **забыть при чем/при ком?** *забыть при декларировании, при зачтении…* (3)
  - **забыть по чему/по кому?** *забыть по рассеянности, по невнимательности* (2)
  - **забыть из-за чего/из-за кого/почему?** *забыть из-за волнения, из-за спешки* (2)
  - **забыть за чем/за кем?** *забыть за давностью* (1),
  - **забыть от чего/от кого/откуда?** *забыть от волнения* (1)
- С какими существительными сочетаются морфемные паронимы **вероятный** Vs. **вероятностный**?

  вероятный определяет существительные *адрес, альтернатива, вариант, версия, встреча…*, а

  вероятностный — *автомат, алгоритм, анализ, анализатор, аспекты…*, и пересечение этих множеств ничтожно мало.
- С какими существительными сочетаются омографы **доменный1** Vs. **доменный2**?

  доменный1 определяет существительные *адрес, аукцион, бизнес, границы, зона, имена…*, а

  доменный2 — *газы, кокс, конструкция, мастера, печь…*, и пересечение этих множеств пусто.
- С какими существительными сочетаются квазиомографы **личный** Vs. **личной**?

  личный определяет существительные *автомашина, адъютант, амбиции, антипатии, архив…*, а

  личной — *карман, крем, напильник, нашивки, полотенце, салфетка…*, и пересечение этих множеств незначительно.
- Что означают и что определяют прилагательные *ретроактивный, проактивный, адвалорный, халяльный…*? На это отвечают их синонимы и связанные существительные.

Диалоговые запросы для уже продвинутого в русском языке иностранца включают все средства, предложенные русскоязычному пользователю, и многое иное:

- Среди сведений по орфографии и морфологии слов можно, например, узнать, что ***Христос*** склоняется особо.
- Можно увидеть сферы применения синонимов *малый* и *маленький*. Так, равно допустимы *малые дети* и *маленькие дети*, но возможны только *малый бизнес* и *маленькие апельсины*.
- При обращении через английский словарь в виде глагола ***pay*** будут получены русские глаголы ***обращать, обратить, окупать, окупить, оплатить, оплачивать, платить, уделить, уделять, уплатить, уплачивать***, и далее можно справиться о любом из них.

Среди недиалоговых приложений КроссЛексики отметим в первую очередь:

- **Облегчение парсинга.** В предложении ищутся все возможные словосочетания, имеющиеся в КроссЛексике, и чем больше обнаружено таких словосочетаний в данном варианте разбора предложения, тем вероятнее этот вариант.
- **Разрешение неоднозначности слово.** Ищутся словосочетания и семантические связи для отдельных омонимов, и выбирается омоним, для которого в контексте найдено наибольшее число синтаксически и семантически сочетающихся соседей.
- **Стеганография и стеганализ.** Сочетания и синонимы слов, встреченные в тексте, используются для регулируемой замены одних синонимов другими, так чтобы в этих заменах закодировать стороннюю информацию, тем самым тайно передающуются несущим текстом без изменения его смысла.
- **Идиоматичный перевод английских словосочетаний**. Например, в ответ на введенное *strong woman* словарь выдает *крепкая баба, сильная женщина…*
- **Информационный поиск**. Предполагается автоматически обогащать запрос не только семантически связанными словами, но и словами, формирующими высокочастотные словосочетания со словами запроса.

4. Источники и метод пополнения словаря, покрытие им текстов

Базовым методом подбора материала был ручной. На момент начала разработки ни корпусов русских текстов, ни интернетовских поисковиков, ни идей работы с ними просто не существовало. Однако на каждом этапе уже наличествующая версия словаря выявляла необходимость очередных его пополнений.

Основными источниками словосочетаний явились:

- Двуязычные словари (особо отметим словарь под ред. Ю.Д. Апресяна и русско-испанский словарь Г.Я. Туровера и Х. Ногейры);
- Академический четырехтомный словарь русского языка;
- Множество специализированных словарей по экономике, бизнесу, электронике, вычислительной технике и др.;
- Наблюдаемый шесть лет поток новостей, политических и научных статей портала *газета.ру*;
- Многочисленные справки по комбинаторике слов в Яндексе и Гуггле;
- Систематические сканирования текстов в рекламных буклетах, объявлениях по ремонту и строительству, в журналах для автомобилистов, в гламурной журналистике, в спаме.

Из Национального корпуса русского языка не было взято ничего. Он появился слишком поздно и вначале был очень небольшим, а для свободного поиска со статистическими оценками результатов недоступен и сейчас.

Методы автоматического извлечения коллокаций из корпусов и Интернета начали разрабатываться лишь в последние годы [8, 9]. Но они прямо не приложимы к высокофлективному языку, да и не дали новых словарей английских коллокаций.

Для проверки через интернет, стоит ли включать в КроссЛексику данное словосочетание, полученное откуда угодно, была предложена количественная мера [10], успешно применяемая к новым пополнениям КроссЛексики.

В части оценок покрытия КроссЛексикой отдельных слов и словосочетаний автоматических средств пока не разработано, но проводились ручные эксперименты. Если исключить названия организаций и географических объектов и личные имена, то покрытие знаменательных слов уже несколько лет близко к 100%. В части же покрытия словосочетаний произошедшее за последние 12 лет увеличение их числа в КроссЛексике в три раза привело к сдвигу примерно с 60% до 75%. Однако последняя цифра резко колеблется от текста к тексту, и требуются новые массовые независимые оценки. Двумерный закон Ципфа неумолим, и гарантированное покрытие хотя бы 80% словосочетаний потребует новых колоссальных усилий. Стопроцентное же покрытие едва ли возможно из-за авторской свободы употребить *ответственный за шишки, минюстовский блин комом* и под.

5. Заключение

Предложен новый словарный ресурс — комбинаторный словарь КроссЛексика, по объему и структуре не имеющий аналогов ни для одного языка. Он оставляет далеко позади единственный для английского языка словарь коллокаций [7] и существенно превышает русские словари [2] и [4].

При высочайшем покрытии лексики и сравнительно высоком покрытии словосочетаний, а также при простом доступе КроссЛексика предназначается для широкого круга пользователей.

Литература

1. *Квеселевич Д. А.* Толковый словарь ненормативной лексики русского языка. Москва: Астрель АСТ, 2005, 1022 стр.
2. *Комплексный* словарь русского языка. Под ред. А.Н. Тихонова. Москва: Русский язык медиа, 2007, 1230 стр.
3. *Крысин Л. П.* Толковый словарь иноязычных слов. Москва: Эксмо, 2008, 942 стр.
4. *Словарь* сочетаемости слов русского языка. Под ред. П. Н. Денисова и В. В. Морковкина. Москва: Русский язык, 1983, 686 стр.
5. *Толковый* словарь русского языка начала XXI века. Под ред. Г.Н. Скляревской. Москва: Эксмо, 2007, 1132 стр.
6. *Mel'čuk, I.* Phrasemes in Language and Phraseology in Linguistics. In: M. Everaert et al. (Eds.) Idioms: Structural and Psychological Perspectives. Lawrence Erlbaum Associates Publ., Hillsdale, NJ / Hove, UK, 1995, p. 169–252.
7. *Oxford* Collocations Dictionary for Students of English. Oxford University Press, 2003.
8. *Lin, Dekang.* Extracting Collocations from Text Corpora. First Workshop on Computational Terminology, Montreal, Canada, August, 1998.
9. *Kilgarriff, A., P. Rychlý, P. Smrz, D. Tugwell.* The Sketch Engine. Practical Lexicography: A Reader, Oxford University Press, UK, 2008, p. 297–306.
10. *Bolshakov, I. A., E. I. Bolshakova, A. P. Kotlyarov, A. Gelbukh.* Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 9th Intern. Conf. on Computational Linguistics CICLing-2008, Haifa, Israel. LNCS 3878, Springer, 2008, p. 95–116.