

# БОЛЬШОЙ ЭЛЕКТРОННЫЙ СЛОВАРЬ КАК ПОЛИТЕМАТИЧЕСКИЙ СПРАВОЧНИК И ФОРМИРОВАТЕЛЬ ЗАПРОСОВ К ИНТЕРНЕТУ

**И. А. Большаков** (bolshakov34@mail.ru)

Независимый исследователь, Москва, Россия

**А. Ф. Гельбух** (gelbukh@gelbukh.com)

Национальный политехнический институт, Мехико, Мексика  
и Университет Васеда, Токио, Япония

Описывается большой электронный словарь, содержащий как фундаментальные сведения о русском языке (грамматические свойства слов, их комбинаторику, семантические и паронимические связи слов), так и обширные энциклопедические сведения о географических объектах, известных персоналиях, организациях и артефактах. В словаре содержатся технические термины, базовые понятия точных и гуманитарных наук, бизнеса и экономики. Словарь позволяет быстро формировать и прямо направлять в Интернет запросы медицинского, коммерческого, туристического и др. характера.

**Ключевые слова:** большой электронный словарь, электронный словарь, справочник, интернет-запросы, запросы к Интернету

## A LARGE ELECTRONIC DICTIONARY AS A POLYTHEMATIC GUIDE AND SHAPER OF QUERIES TO THE WEB

**I. A. Bol'shakov** (bolshakov34@mail.ru)

Independent researcher, Moscow, Russian Federation

**A. F. Gel'bukh** (gelbukh@gelbukh.com)

National Polytechnic Institute, México

A large Russian electronic dictionary is presented. It contains both fundamental information on the Russian language (grammatical and combinatory properties of words, semantic and paronymic relations between words) and ample encyclopedic information on geographical objects, famous people,

organizations, and artifacts. The dictionary includes technical terms and basic concepts of science, humanities, business, and economy. Among its applications is the possibility to form queries for Internet search engines on medicine, commerce, tourism, and other topics.

**Key words:** large electronic dictionary, electronic dictionary, guide, Internet search, queries

## 1. Введение

Лингвистика оперирует знаниями двух типов: чисто лингвистическими (о грамматике слов и их комбинаторике) и энциклопедическими (о внешнем мире). Общеобразовательные словари в основном содержат знания первого типа, хотя часто включают перечни географических имен и персоналий, т. е. энциклопедических объектов. Все более необходимые в обыденной жизни энциклопедические знания помещают в специализированные словари и энциклопедии.

Для экономии поисковых усилий возникает желание иметь электронный словарь, содержащий как можно больший объем лингвистических знаний вместе с необходимым минимумом энциклопедических сведений.

Современному человеку всегда требуется и информация о текущем состоянии внешнего мира: *Где купить X? Каков срок действия документа Y? Как лечить болезнь Z? Каковы достопримечательности страны W?* Традиционно все это отражается в прессе, радио- и телевизионных передачах, а теперь еще и в Интернете.

Казалось бы, Интернет покрывает сейчас информационные потребности любого пользователя. Но поиск в сети порой требует определенных усилий: нужно уметь разумно составить запрос и быстро выбрать нужный сайт среди появившихся на экране компьютера сниппетов. Поэтому хочется иметь лингвистическое средство, которое сочетало бы свойства универсального словаря-справочника указанного выше типа с возможностью формирования релевантного запроса к Интернету.

Данная работа содержит описание КроссЛексики (КЛ) — большого электронного словаря русского языка, адекватно отвечающего поставленным выше требованиям: выдавать любые лингвистические и многие энциклопедические справки, а также формировать типовые запросы к Интернету. Предыдущие публикации о КЛ [2] и [4] представляли этот словарь как чисто лингвистический ресурс, и содержащиеся в нем энциклопедические знания не упоминались. Новая функция КЛ как формирователя запросов к Интернету оставалась неосознанной и практически не реализованной. В данной статье мы восполняем этот пробел.

Наше изложение имеет целью описать:

- Структуру КЛ в целом — для облегчения понимания дальнейшего;
- Части структуры КЛ, содержащие особо богатые лингвистические сведения;
- КЛ как словарь-справочник энциклопедического характера;
- Одноступенчатое формирование запроса к Интернету: через словник КЛ;
- Двухступенчатое формирование запроса к Интернету: через элементы выдачи КЛ;

- Преимущества и недостатки КЛ как формирователя запросов в сравнении с непосредственным обращением в Google.  
Указываемые ниже статистические параметры соответствуют марту 2011 г.

## 2. Краткий обзор КроссЛексики

В основе КЛ лежит квадратная матрица {словник  $\times$  словник}, где словник — это вектор  $\{t_1, t_2, \dots, t_n\}$  из титулов, являющихся отдельными словами или коллокациями (рис. 1). Размер словника  $n$  по состоянию на март 2011 г. превысил 250 тыс. Элемент  $D_{ij}$  этой матрицы является дескриптором односторонней связи между  $t_i$  и  $t_j$ . Связь бывает синтагматической (титулы образуют коллокацию), семантической (титулы связаны смысловым сходством типа синонимии, антонимии, гипонимии / гиперонимии и др.) или паронимической (титулы связаны буквенным или морфемным сходством). Например, титулы  $t_i = \text{борьба}$  и  $t_j = \text{суетерия}$ , образующие коллокацию *борьба против суетерий*, имеют  $D_{ij} = \{t_i \text{УПРАВЛЯЕТ 'против' } t_j\}$ , а титулы  $t_i = \text{диплом}$  и  $t_j = \text{документ}$  имеют  $D_{ij} = \{t_i \text{ГИПОНИМ } t_j\}$ .

Запрос к КЛ в виде титула  $t_i$  (на рис. 1 возможные запросы представлены словником в крайнем левом столбце) ведет к выдаче всех тех титулов, связь которых с  $t_i$  зафиксирована в матрице. Совокупность потенциально выдаваемых титулов дается словником в верхней строке. Если связь является синтагматической, то в выдаче автоматически формируется коллокация в ее грамматически правильной форме, учитывающей род, число, падеж и одушевленность русских существительных и прилагательных.

Рассматриваемая гигантская матрица очень разрежена: непустым оказывается примерно один дескриптор на 7600. Однако уже выявлено 6,93 миллионов непустых ее элементов. Вот важные особенности матрицы:

- Титул любой из четырех частей речи может быть как одиночным словом, так и многословным оборотом. При этом знаменательные составляющие оборота входят в словник и автономно, а декомпозиция может быть многоступенчатой: (((судебные) (прения)) и (((последнее) (слово)) (подсудимого))).
- Если непуст дескриптор  $D_{ij}$ , то непуст и обратный ему  $D_{ji}$ . Например, для  $t_i = \text{диплом}$  и  $t_j = \text{документ}$  имеем  $D_{ji} = \{t_i \text{ГИПЕРОНИМ } t_j\}$ .

В КЛ встроены входы в Google, Яндекс и Национальный корпус русского языка. Два поисковика являются самыми мощными в Рунете и обслуживают миллионы пользователей. НКРЯ, крупнейший корпус частично размеченных русских текстов, может служить источником примеров словоупотреблений в контексте.

## 3. КроссЛексика как лингвистический справочник

Хотя словарь КЛ является единым целым, именованные секции его выдачи, рассматриваемые совместно, можно называть подсловарями. Ряд подсловарей КЛ содержит особо богатую лингвистическую информацию.

- Подсловарь **Коллокаций** покрывает все типы русских коллокаций и включает их в количестве 1,96 млн. (3,92 млн. односторонних связей), см. [2, 4].
- Подсловарь **Модели управления** характеризует более 20 тыс. управляющих глаголов и 15 тыс. существительных. Модели управления имеют также многие прилагательные и наречия.
- Подсловарь **Синонимов** уникален по объему: 115 тыс. синонимов разбиты на 21 тыс. групп, и общее число связей превышает 1,2 млн. (ср. с 600 тыс. связей в [1]).
- Подсловарь **Ассоциаций** не имеет прецедентов. Он создан на основе сочиненных пар, образующих частые запросы к Интернету или многократно представленных на интернет-сайтах [3]. На данный момент найдено 58,1 тыс. ассоциаций для 15,3 тыс. одно- и многословных титулов, в среднем 3,8 ассоциаций на титул.
- Подсловарь **Морфопарадигм** дает все возможные формы для каждого изменяемого титула. Падежные формы многословных именных титулов могут включать до шести частей, из них до пяти изменяемых, например, *десять заповедей и семь смертных грехов*.
- Подсловарь **Семантических деривативов** составлен из подсекций титулов, выражающий одно и то же понятие четырьмя главными частями речи: существительными, глаголами, прилагательными или наречиями. Наиболее просты здесь случаи, когда семантическая деривация в основном осуществляется морфологическими средствами:

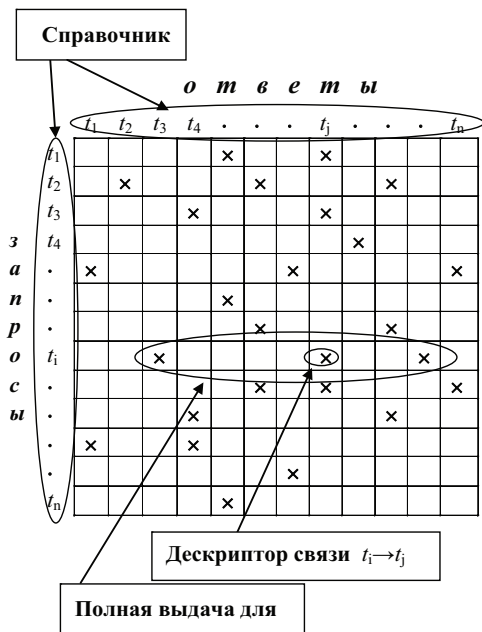


Рис. 1. КроссЛексика как матрица связей

**Сущ** *владение, владения, владелец, владельцы, собственник, собственники, собственность*

**Глаг** *владеть, завладеть, иметь в собственности, находиться во владении, являться собственностью*

**Прил** *владельчий, владевший, завладевший, находящийся во владении, собственный, являющийся собственностью*

**Прич** *будучи собственностью, владея, в качестве собственности, в собственности, в собственность, во владении, завладев, как собственник, овладев*

В подсекции существительных встречаются два числа одного существительного, а в подсекции глаголов — разные виды одного русского глагола.

- Подсловарь **Гипонимов и гиперонимов** является обширной полииерархией понятий. Например, *Франция* имеет два гиперонима — *европейская страна* и *средиземноморская страна*, а *цветы* имеет гиперонимами *предметы быта* и *растения*. Наличие нескольких уровней иерархии позволяет транзитивно переносить свойства гиперонимов на гипонимы несколькими уровнями ниже, что в КЛ активно используется.
- Поскольку однозначно принятого соотношения «часть Vs. целое» не существует, понятия **меронима** и **холонима** у нас многозначны. Так, существительное множественного числа считается холонимом единственного числа. Меронимом считается часть физического объекта, например, *нога* по отношению к *телу* или *рулевое колесо* по отношению к *автомобиль*. Для неисчисляемых объектов используется понятие **квантификатора**. Так, для *вода* квантификаторами являются *капля* / *стакан* / *бочка* / *цистерна воды*, а для *гнев* — *приступ гнева*. Перенос свойств от холонимов к меронимам не предусмотрен.

#### 4. КЛ как справочник энциклопедического характера

КЛ покрывает следующие тематические области:

- Экономика и бизнес;
- Общественно-политическая тематика (политика, социология);
- Техника и технологии (радиоэлектроника, компьютеры, программирование, Интернет, бытовая техника, автомобили, строительство);
- Точные и естественные науки (математика, физика, химия, биология, география, география);
- Медицина (не только бытовая);
- Гуманитарные науки (лингвистика, психология, философия, история), искусство, религия;
- Бытовой язык (включая бранную лексику, но без нецензурной).

Термины и имена из указанных областей и представляют собой энциклопедическую информацию. Остановимся на этом подробнее.

В подсловаре **Семантических деривативов** (далее ПССД) имеется несколько десятков наборов, характеризующих наиболее известные страны. В подразделе

существительных здесь даются наименования государства и наций, ее представитель мужского и женского пола, официального языка, столицы, титула главы государства, единицы административного деления, денежной единицы, официальной или ведущей религии (если таковая существует). Например, для титула *Франция* даются *Париж, француз, французы, француженка, француженки, французский язык, президент Франции, провинция Франции, евро, франк*. Данные здесь же коллокации могут служить отсылками к дополнительной информации внутри КЛ или вне ее: *город Франции, провинция Франции, население Франции, площадь Франции, туры во Францию, отдых во Франции, достопримечательности Франции*.

В ПССД имеется также несколько десятков наборов, характеризующих российские города и включающих наименования их жителей. Не все говорящие по-русски сразу вспомнят, как называются жители Архангельска, Смоленска, Тулы, Курска, Нижнего Новгорода или Пскова, а в отношении жительниц этих городов ситуация еще сложнее, поскольку для ряда городов их называют только описательно.

В подсловаре **Гипонимов и гиперонимов** (далее ПСГГ) содержатся важные географические понятия (моря, океаны, континенты, горы и др.) и их имена. Даются также имена городов России и множества зарубежных стран. Приведены также имена единиц административного деления, например, области или края России, провинции Франции, штата США, Канады, Мексики и Индии, графства Великобритании, воеводства Польши.

В ПСГГ содержатся также наиболее известные персоналии, в основном, из прошлого. Это ученые, композиторы, поэты, политические и общественные деятели, миллиардеры. Здесь же содержатся имена наиболее известных культурных артефактов, как то романов, опер, оперетт, мюзиклов, фильмов, мультфильмов.

В ПСГГ приведены 48 видов цветов, 152 оттенка цвета, 183 специализации заводов, 124 специализации промышленности. Широко представлены предметы быта, включая мебель, электрооборудование и гаджеты.

В подсловаре **Определений** содержится более 200 тыс. научно-технических терминов со структурой «прилагательное ← существительное», где существительное-гипероним служит терминообразующим ядром. Наиболее продуктивны *режим<sub>2</sub>* (514 терминов-гипонимов), *покрытие<sub>1</sub>* (438), *анализ* (423), *препараты* (416), *конструкция<sub>1</sub>* (386), *вещества* (377), *контроль* (364), *детали<sub>2</sub>* (364).

Продемонстрируем извлечение из КЛ энциклопедической информации о Франции (рис. 2). Войдя в КЛ, мы сразу оказываемся в словнике, где набором пяти начальных букв попадаем в строку *Франция*. Нажав *Enter*, получаем на экране соответствующую Франции выдачу. В секции **Синонимы** приведены известные фигуральные названия страны — *галльский петух* и *страна гурманов*. **Гиперонимы** свидетельствует, что Франция и европейская и средиземноморская страна. Ее **Когипонимы** — это многочисленные страны, образующие эти две группы — среди них, например, *Австрия* как страна Европы и *Алжир* как страна Средиземноморья. В секции **Семантические деривативы** особо интересна подсекция существительных. Выбрав *город Франции*, получаем новую выдачу, где **Гипонимы** — длинный список французских городов: *Авивиль, Авиноньон, Авориаз, Адье,...* Выбрав *президент Франции*, получим **Синонимы** в виде перифраз *Николя Саркози, президент Николя Саркози, президент Саркози*.

Выбрав провинция Франция, получим список Аквитания, Бретань, Бургундия, Лангедок, Лотарингия и т.д.

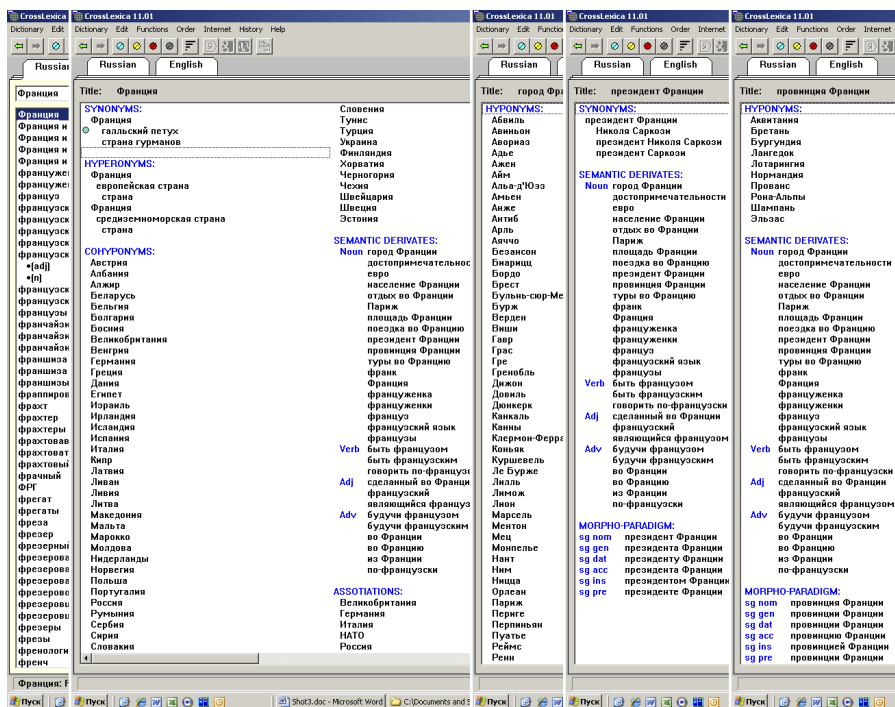


Рис. 2. Поиск в КЛ информации о Франции

## 5. Одноступенчатое формирование запроса к Интернету

Размер словника достигает в настоящее время 250 тыс. единиц. Каждый его титул можно сразу направить в Интернет в качестве запроса. Это одноступенчатое его формирование. Скорее всего, это будут именные титулы, которых в словнике более 90 тыс.

Одним шагом могут быть сформированы запросы типа *боль в горле / желудке / ногах / шее, туры в Египет / Испанию / Италию / Израиль / Турцию / Финляндию*. Особенно интересны для одноступенчатых запросов сочиненные пары типа *простуда и антибиотики, радиация и иммунитет, лечебное питание и климатолечение, герметизация и теплоизоляция, боль и тяжесть в желудке, боль и хруст в коленях, светильники и абажуры / бра / люстры / лампы, курага и диабет / запор / сердце / чернослив*. Таких пар в словнике 34 тыс.

Продemonстрируем одноступенчатое формирование запроса на примере титула *простуда и антибиотики* (рис. 3). Достаточно набрать в словнике шесть первых букв, как на экране появляется искомым титул. Подведя к нему курсор и нажав иконку Google, получим совокупность релевантных сниппетов.

Для диверсификации конечного вида запроса в КЛ предусмотрены четыре опции: 1) отсылка запроса  $Q$  как есть; 2) отсылка « $Q$ » в кавычках; 3) отсылка  $Q$  это, т. е. с постфиксом «это»; 4) отсылка « $Q$  это» в кавычках. Первая опция очень мало ограничивает поиск. Вторая выдает только буквальное вхождение запроса в тексты сайтов. Третья опция ищет определения запрашиваемого понятия (например, *валоризация это*) и, как правило, выдает определение из Википедии. Четвертая опция примерно эквивалентна третьей.

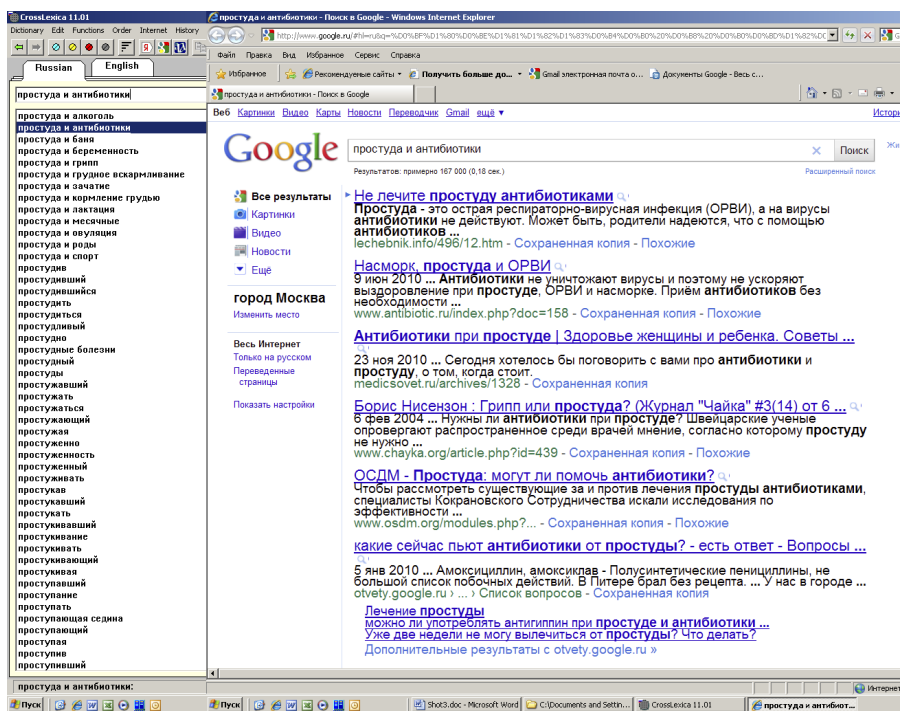


Рис. 3. Одноступенчатый поиск для *простуда* и *антибиотики*

## 6. Двухступенчатое формирование запроса к Интернету

Можно брать в качестве запросов к Интернету и любой элемент выдачи для произвольного титула. Число таких элементов равно суммарному числу связей в КЛ, т. е. 6,93 млн. Это двухступенчатое формирование запроса. Берется титул в словнике, вызывается на экран его выдача, и какой-либо ее элемент отсылается в Интернет. В первую очередь интересно использовать в качестве запроса любую из 1,96 млн. колокаций.

На рис. 4 отражены этапы формирования запроса *купить бинокль*. В окошке Словника вводим начальные буквы слова *купить*. В секции **Модели управления** его выдачи выбираем подсекцию *купить что / кого?*, в ней подводим курсор к строке *бинокль* и нажимаем иконку поисковика. Точно так же



A large electronic dictionary as a polythematic guide and shaper of queries to the web

ищется *лечить артроз / варикоз / гастрит / диатез / холецистит, найти врача / няню / тренера, воспаление лимфоузлов / печени / почек / слизистой / сосудов.*

Для запросов в виде коллокации всегда существуют два эквивалентных двухступенчатых пути в Интернет. Так, запрос *купить бинокль* можно сформировать и через секцию *Управляется\_глаголами* в выдаче для *бинокль*.

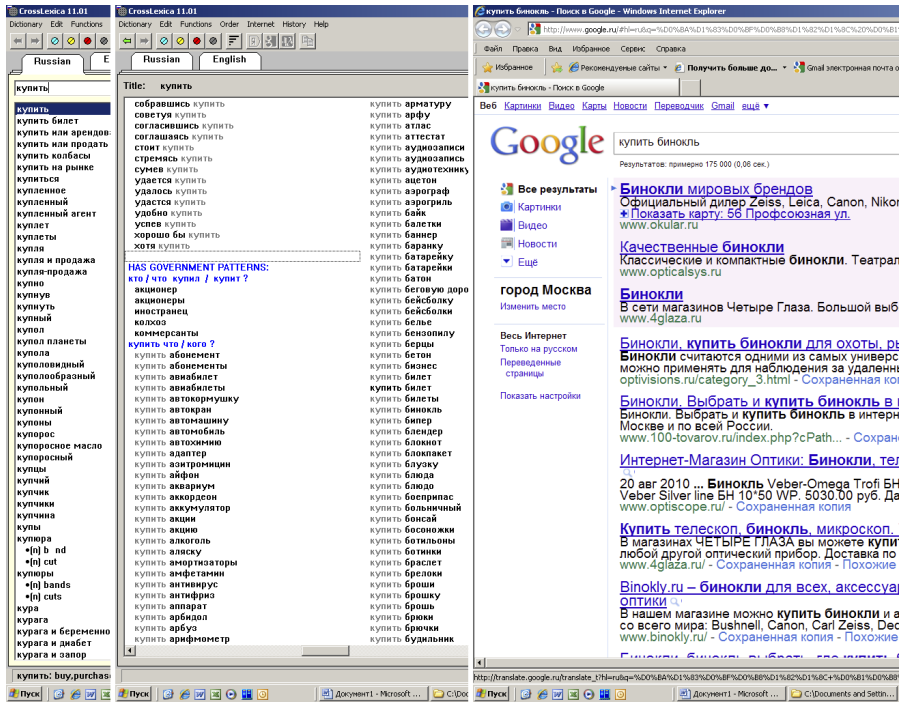


Рис. 4. Двухступенчатый поиск для *купить бинокль*

## 7. Преимущества и недостатки КЛ как формирователя запросов

Поясним преимущества и недостатки КЛ как формирователя запросов к Интернету по сравнению с непосредственным вводом в Google.

Как только пользователь входит в Google, он соединяется с накопленной базой прежних запросов, из которых Google оперативно формирует меню подсказок. Своей начальной частью элементы меню повторяют уже введенную строку и наиболее частые из прежних запросов. Подсказок всегда немного, обычно 4–6. Одна подсказка иногда вкладывается в другую, так что реально элементов в меню оказывается еще меньше.

Ввод существенно зависит еще от того, включен ли режим так называемого «живого поиска». При широкополосном Интернете живой поиск включен

всегда, при узкополосном — когда поисковик недогружен. Если живой поиск отключен, поисковик не ищет нужные сайты, пока пользователь не нажмет *Enter*. При включенном живом поиске Google постоянно ищет сайты и мгновенно выдает их сниппеты на экран даже при самой короткой или абсурдной строке, введенной пользователем.

И в базе запросов, и в основном массиве Google за годы накопилось много неверных начальных цепочек. Поэтому при вводе первых букв запроса возможно появление неграмотных или абсурдных подсказок и сниппетов, соответствующих накопленным в прошлом ошибкам. Это цена, которую приходится платить за статистические методы работы поисковика.

КроссЛексика не пользуется ни базой запросов, ни живым поиском. Фактически, она заменяет базу запросов. Достоинствами ввода запроса через КЛ является то, что:

- В окошке ввода КЛ появляется продолжение введенной цепочки букв в виде существующего титула, ближайшего в словнике по алфавиту. Если пользователь начал вводить абракадабру, поиск в Словнике останавливается там, где пользователь свернул с правильной дороги.
- Меню в КЛ всегда много обширнее, чем у Google. Ввод первых 3–6 букв обычно позволяет видеть в пределах экрана строку, нужную пользователю. Остается только подвести к ней курсор и нажать иконку Google. Это значит, что КЛ позволяет быстрее перейти от буквенного ввода к перемещению курсора вдоль выведенных на экран упорядоченных по алфавиту альтернатив.
- Содержимое экрана КЛ, в отличие от случая живого поиска, не меняется суетливо по мере ввода очередной буквы и не отвлекает пользователя от процесса ввода, ввод проходит более плавно.
- Независимо от того, включен ли живой поиск или нет, при вводе запроса в Google в любой момент может возникнуть период ожидания до нескольких секунд, когда процесс ввода блокируется. Это может раздражать пользователя.

Если говорить о недостатках ввода запросов через КЛ, то ими является следующее:

- КЛ, как и положено словарям, является ресурсом инерционным и избирательным. Если идет речь о событии, продукте или лице, внезапно возникшем в информационном поле, то сведения о них не будут найдены в словаре, и составляя запрос относительно них через КЛ нецелесообразно.
- В КЛ мало конкретных названий артефактов, даже устоявшихся за годы (например, марок технических изделий). Поэтому запрос, включающий торговую марку, нужно направлять поисковику напрямую.
- В КЛ отсутствует подсказка в виде более вероятного варианта, очень похожего на введенный пользователем запрос. Иногда подсказка помогает, но для опытного пользователя, не допускающего тривиальных ошибок, может явиться «медвежьей услугой»: он ищет одно, а ему подсвывают другое, забывая начальную часть экрана ненужным материалом.

## 8. Заключение

Современный пользователь электронных словарей скорее предпочтет единый словарь, содержащий как всесторонние знания о языке, так и спектр энциклопедических познаний. Предлагаемый словарь в существенной мере является таковым. В нем хранятся имена, неизменные за годы и десятилетия. В целом структура, содержание и объем хранящейся в КроссЛексике информации не имеют аналога в электронном мире в виде единого продукта.

Однако для получения актуальной информации, изменяющейся в течение недель, дней или часов, необходимо обращаться в Интернет. При желании пользователя иметь в своем компьютере единый информационный интерфейс можно потребовать от электронного словаря еще и умения помочь быстро и удобно составить запрос к Интернету по темам, которые сам словарь отражать не может из-за их изменчивости. КроссЛексика предоставляет средства формирования и отсылки запросов по самой широкой тематике, в первую очередь — медицинской, коммерческой или туристической.

Пополнение и совершенствование КроссЛексики продолжается.

Работа выполнена при частичной поддержке правительства Мексики (CONACYT 50206-H, SIP-IPN 20113295, SNI) и Индии (DST India).

## References

1. *Aleksandrova Z. E.* 2005. Russian Synonyms Dictionary [SLovar' Sinonimov Russkogo Iazyka].
2. *Bol'shakov I. A.* 2009. KrossLeksika: Large Electronic Dictionary of Russian Words Combinations and Connections [KrossLeksika: Bol'shoi Elektronnyi Slovar' Sochetanii I Smyslovykh Sviazei Russkikh Slov] . Komp'uternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15) : 45–50.
3. *Bol'shakov I. A., Bol'shakova E. I., Gel'bukh A. F.* 2010. Associative Net of Concepts forming Queries to the Web [Assotsiativnaia Set' Poniatii, Obrazuiushchikh Zaprosy k Internetu]. Komp'uternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"), 9 (16): 55–61.
4. *Bol'shakov, I. A.* 2004. Getting One's First Million Collocations. Computational Linguistics and Intelligent Text Processing. Proc. 5th Intern. Conf. on Computational Linguistics CICLing-2004 : 229–245.