УДК 81

Rus. | Eng.

Эволюция автономного словаря ASIS®
к интерактивному общедоступному веб-словарю

Куянов Юрий Владимирович
старший научный сотрудник
НИЦ «Курчатовский институт»
ФГБУ ИФВЭ, г. Протвино
Тел. (4967)-71-39-67
Эл. почта: Yu.Kuyanov@ihep.ru
Сайт:
http://hermes.ihep.su:8001/compas/kuyanov/

Тришин Виталий Николаевич
к.ф.-м.н.,
Председатель Совета директоров
ООО «ОКП», г. Москва
Тел. (495)-222- 22-58
Эл. почта: mail@trishin.ru
Сайт: www.trishin.ru

(журнал «Научное обозрение: гуманитарные исследования», № 10, 2016 г.)

Аннотация

В статье представлен самый БОЛЬШОЙ РУССКИЙ СЛОВАРЬ-СПРАВОЧНИК СИНОНИМОВ (близких по смыслу слов), автор которого — ТРИШИН Виталий Николаевич напряжённо работал над его созданием примерно с 1992 года. Оценки активов таких крупных предприятий, как АвтоВАЗ, Норильский никель, Сургутнефтегаз и др., которые Тришин выполнял один с помощью ASIS® и входных данных от нескольких ассистентов, вполне можно отнести к областям с интенсивным использованием данных. Впоследствии, когда задачи адекватной оценки активов крупных предприятий потеряли в значительной степени свою актуальность, дальнейшее развитие Большого русского словаря-справочника синонимов приобрело самостоятельное значение. В 2016 году Куяновым Ю. В. реализована сетевая версия словаря trishin.net, описание которой дано в данной совместной статье. В сетевой версии достигнута высочайшая скорость отклика на интерактивные запросы пользователей. Словарь-справочник является попыткой продолжения словаря В. И. Даля на современной живой бесцензурной языковой основе (кроме обсценной лексики) с внесением в словарь также специализированных слов русского языка практически из всех отраслей знаний. Данный словарь-справочник является и частично толковым, так как практически все толкования из толковых (и специализированных) словарей в нём имеются, и они связаны синонимическими связями с толкуемыми словами. Включены в словарь крылатые выражения, пословицы, поговорки и т.п. Краткое описание других онлайн-словарей приведено в [6].

Ключевые слова: Большой русский синонимический онлайн-словарь Виталия Тришина, компьютерная обработка текстов, интеллектуальный поиск информации

The evolution from standalone dictionary ASIS® to interactive public web-dictionary trishin.net

© Yu. V. Kuyanov
NRC "Kurchatov Institute"
FSFI IHEP, Protvino

Yu.Kuyanov@ihep.ru

© V. N. Trishin
candidate of Phys. and Math. Sciences,
Chairman of the Board of Directors
"OKP" Ltd., Moscow
mail@trishin.ru

Abstract

The article presents the largest Russian dictionary of synonyms (similar meaning words), whose author is Trishin V.N. since 1992. The appraises of assets of such large-scale concerns as AvtoVAZ, Norilsky Nikel, SurgutNefteGaz etc, that Trishin made alone with the aid of ASIS® and data entries from some assistants, may indeed be treated as the Data Intensive Domain. Later when the jobs of appraiser support of adequate assets of large-scale concerns suddenly lost their topicality, the development of the large Russian dictionary of synonyms acquired its own meaning. In 2016 Kuyanov Yu.V. has implemented the network version of the dictionary trishin.net as narrated in this joint article. In the online version the highest speed of response to interactive user requests is achieved. This is an attempt to continue the V.I. Dahl tradition on the modern living uncensored language basis with words from virtually all branches of knowledge added.

Оглавление

1 Предыстория системы ASIS® и её словаря

Система помощи оценщику и аудитору ASIS® (Appraiser Support Info System) предназначена для обработки больших массивов информации и позволяет решить разнообразные задачи в области анализа, оценки и аудита крупных имущественных

комплексов, повышая производительность труда оценщика в десятки раз по сравнению с традиционными методами работы.

Система ASIS® написана на языке Visual FoxPro и включает в себя элементы искусственного интеллекта. В ней реализованы практически все существующие методики и базовые расчётные данные для оценки объектов как движимого, так и недвижимого имуществ затратным подходом. Задачи, решаемые в рамках этой системы, описаны в 11 статьях, первая из которых была напечатана ещё в 1995 году [1], а наиболее полно техническая реализация этих задач приведена в статье [2].

Система ASIS® защищена авторским правом (свидетельство о государственной регистрации № 960056, 1996 г., соавторы к.ф.-м.н. Тришин В. Н. и к.т.н. Шатров М. В.). Работа над системой ASIS® была начата ещё в 1990 году в институте «Информэлектро» в рамках договора под № 1 Минимущества с институтом «Информэлектро» (где Тришин В. Н. работал зав. отделом «Экспертные системы») по созданию Базы данных приватизированных предприятий России совместно с Госкомстатом РСФСР [3].

По программе приватизации (1991 г.) первого руководителя Минимущества Малея М. Д. (ранее — директора института «Информэлектро»), приватизация предприятий должна была производиться аналогично приватизации в восточноевропейских странах с помощью именных приватизационных чеков и по справедливой, объективной стоимости. После прихода к руководству в Минимущество команды «молодых реформаторов» в ноябре 1991 года государственное финансирование работ «Информэлектро» для Минимущества было прекращено, так как был выбран курс на ваучерную приватизацию по заниженной на 2–3 порядка стоимости предприятий и выдачей льготных государственных кредитов для «своих» приватизаторов.

В 1992 году разработчики системы ASIS® Тришин В.Н. и Шатров М.В. перешли на работу в компанию «Эрнст и Янг», где занимались практической оценкой предприятий и доработкой программного комплекса. В рамках этих работ была создана программа по автоматическому переводу инвентарных ведомостей основных средств (здания, сооружения, машины и оборудование, транспорт и т.д.) с русского языка на английский для иностранных оценщиков, а также синонимический словарь технических терминов для лучшей идентификации оцениваемых объектов (для этого использовались не только сами слова, но и их основы вместе с кодами 19 эталонных слов, по которым происходит спряжение глаголов, склонение существительных, прилагательных и пр. в русском языке). В дальнейшем этот словарь был выделен в отдельный комплекс в конце 90-х годов на базе Visual FoxPro, его словарная база была существенно расширена и словарь был помещён в Интернет в незашифрованном виде, начиная с 2000 года. Словарная база этого словаря примерно на 324 тыс. записей была «заимствована» вместе с опечатками многими анонимными авторами различных «самых полных» словарей. Поэтому начиная с 2010 года синонимические связи словаря в свободном доступе стали шифроваться. Словарь защищён авторским правом (правообладатель Тришин В. Н., свидетельство о государственной регистрации № 2013616013, 2013 года). В настоящее время эта реализация словаря используется Тришиным В.Н. в качестве рабочего инструмента по пополнению словарной базы и синонимических связей. А электронная «укороченная» версия 7.0 на 526 тыс. слов и словосочетаний с 1949 тыс. зашифрованными синонимическими связями размещена на сайте trishin.ru в качестве демонстрационного варианта.

В полном варианте словаря-справочника на апрель 2016 г. имеется уже более 580 тыс. слов и 2129 тысяч синонимических связей. Работа над пополнением словаря продолжается, и эта работа должна завершиться, в основном, в 2017 году цифрой приблизительно в 610 тыс. записей.

В 2011–2013 годах к.т.н. Ивановым С. К. была реализована и размещена на сайте trishin.ru электронная версия словаря на 480 тыс. слов с зашифрованными

синонимическими связями в более современной программной оболочке (написана на C# в среде разработки Microsoft Visual Studio 2008) и с реализацией интерфейса и инструкций по использованию на 4-х иностранных языках. Система позволяет выполнять практически любые запросы, в том числе и задаваемые в телешоу «Поле чудес». Эта версия программы могла бы стать электронным словарём русского языка для иностранцев, однако на письмо с описанием её возможностей в Комиссию по русскому языку при Президенте РФ (а расширение влияния русского языка за рубежом — одна из задач Комиссии) ответ получен не был. Поэтому работы по данной версии словаря были приостановлены.

2 Что такое синоним и синонимическая связь

Если тебе говорят: «У Вас написано с ошибкой!» — ответствуй: «Так всегда выглядит в моём написании» (Козьма Прутков).

«СИНОНИМЫ (от греческого synonymos — одноименный), слова, различные по звучанию, но тождественные или близкие по смыслу, а также синтаксические и грамматические конструкции, совпадающие по значению. Синонимы бывают полные ("языкознание" — "языковедение") и частичные ("дорога" — "путь")» (*Современная энциклопедия, 2000.*).

Отметим, что степень близости по смыслу является качественной категорией, поэтому мнения о том, являются ли конкретные слова (словосочетания) синонимами или нет, у разных людей могут не совпадать.

Учитывая, в первую очередь, прикладную нацеленность словаря-справочника Тришина В. Н. для компьютерной обработки текстов, мы используем более общее понятие «синонимическая связь», а именно, если в какой-либо фразе одно слово (словосочетание) заменить другим и, хотя бы один смысл фразы сохраняется, то мы говорим, что эти слова (словосочетания) синонимически связаны. Поэтому, например, названия 5656 минералов (абелсонит, абернатит, абихит и т.д.) в словаре синонимически связаны с понятием «минерал». Это даёт возможность поиска нужного слова в словаре по обобщённому понятию, особенно когда пользователь забыл название нужного ему слова, например, запрос (в полной версии 7.1) по слову "рыба" выдаст в алфавитном порядке список из 873 названий различных рыб, по слову "растение" — 4845, "песня" — 168, "певец" — 122, "бард" — 48, "музыкант" — 164, "поэт" — 88, "графоман" — 42, "девушка" — 138, "взрывчатка" — 234, "фермент" — 576 и т.д. То есть словарь является и справочником практически по всем отраслям знаний. Данный словарь-справочник является и частично толковым, так как практически все толкования из толковых (и специализированных) словарей в нём имеются, и они связаны синонимическими связями с толкуемыми словами. Словосочетаний (толкований, фразеологизмов, пословиц, крылатых выражений и т.д.) в словаре порядка 19 процентов.

Предлагаемый словарь имеет несколько особенностей.

1. Отсутствие разделения омонимов, что приводит к тому, что в одной группе синонимов к слову, скажем, «стан» оказываются слова «лагерь», «блюминг» и «талия». Для образованного носителя языка это не представляет сложности, но для иностранца, желающего углубиться в недра русского языка, отсутствие разделения омонимов будет, конечно, представлять сложности. Это несколько ограничивает возможности применения текущей версии словаря в системах компьютерной (автоматической) обработки русских текстов.

Для того же «стана» правильно было бы иметь несколько словарных статей, например

Стан 1 => блюминг, трубопрокатник, листопрокатник, ...
Стан 2 => становище, стойбище, станица, бивуак, лагерь, ...
Стан 3 => торс, талия, поясница, ...

В будущих версиях словаря, возможно, такое разделение будет предусмотрено, и словарь будет состоять из синонимических групп, связанных с некоторыми «понятиями», для которых останется или дать определения из толковых словарей, или выбрать среди синонимов группы так называемую «точку входа», то есть слово, наиболее точно (строго) обозначающее понятие, объединяющее данные синонимы.

2.Отсутствие разделения синонимических и родо-видовых связей. Родо-видовые (цело-частные) отношения — это, вообще говоря, предмет тезаурусов. Синонимическое отношение — это отношение эквивалентности, которое разбивает все слова на классы эквивалентности. А родо-видовое отношение — отношение частичного порядка. В идеале должна быть вкладка для синонимов, для родительских терминов и для дочерних. Такое разделение также планируется произвести в программе словаря после завершения работы над его пополнением.

3 Автономные инструменты

«Рабочим станком» по созданию словаря-справочника является версия словаря 7.1 написанная на Visual FoxPro. При пополнении словарной базы новым словом (словосочетанием) требуется выбрать часть речи и эталонное слово из предлагаемых системой 19 слов, согласно которому склоняется (спрягается) вводимое слово. Тем самым появляется возможность встраивания словарной базы словаря-справочника в различные поисковые системы и анализаторы русских текстов. В системе ASIS® это позволило организовать полуавтоматический поиск близких по смыслу наименований оцениваемых инвентарных единиц крупных предприятий в базах аналогов машин и оборудования системы ASIS® и с последующей автоматической оценкой этих единиц, а также присвоение инвентарным единицам кодов согласно государственному классификатору основных фондов (ОКОФ). На крупных предприятиях может быть несколько сот тысяч (и даже миллионов) оцениваемых инвентарных единиц, в базе аналогов системы ASIS® имеется порядка 200 тыс. аналогов, в классификаторе ОКОФ свыше 11 тыс. групп машин и оборудования, транспорта и пр. И без синонимического словаря, кодирования смысла наименований инвентарных единиц, аналогов, названий групп в классификаторе ОКОФ оценка затратным подходом имущественного комплекса крупного предприятия представляла бы собой исключительно трудозатратную задачу. И система ASIS® «понимает», к примеру, что битумоварка из оцениваемого оборудования является, возможно, котлом для варки битума из базы аналогов.

Если для введённого в словарь-справочник нового слова (словосочетания) имелись слова (словосочетания), которые были признаны «синонимами» либо Тришиным, либо различными авторами синонимических (или специализированных) словарей, то между этими словами (словосочетаниями) устанавливалась синонимическая связь. В противном случае вводилось в словарь отсутствующее слово-синоним с последующей организацией синонимических связей. Все эти операции занимали считанные секунды практической работы автора словаря.

Кроме того, при создании словаря использовались несколько утилит:

А) программа по назначению обратных ссылок для заведённых новых синонимических связей;

Б) программа поиска и удаления задвоенных синонимов в словарной базе (в определённых ситуациях программа словаря иногда пропускала «задвоенность»);

В) программа шифрации синонимических связей.

Г) программа отбора слов в произвольном тексте (обычно из Интернета), отсутствующих в словаре. Найденное «новое» слово в выбранном тексте анализировалось на основе других источников и по ним принималось Тришиным решение включать слово в словарь или нет. Как правило, слова-кандидаты из анализируемых текстов отвергались, так как они были опечатками уже известных слов из словаря.

4 Источники пополнения базы слов и синонимических связей

При составлении словаря использованы изданные словари: орфографические, синонимические, фразеологических синонимов, толковые, начиная со «Словаря русского языка XVIII века» и «Толкового словаря живого великорусского языка» В. И. Даля (полностью в версии словаря-справочника 7.1), (т. е. примерно за последние триста лет), церковно-славянских слов, иностранных слов, неузуальных слов, «Словарь-справочник по материалам прессы и литературы 90-х годов XX века», арго (кроме заведомо ненормативной лексики), а также специализированные словари практически по всем отраслям знаний. Кроме того, для пополнения словаря активно использовались газеты, журналы, Интернет, словари поисковиков Yandex и Google путём анализа выдачи их автоподсказчиков слов (с проверкой по другим источникам), так как даже в самых крупных печатных словарях отсутствуют десятки тысяч широко распространённых в быту и прессе современных слов, например, паранормальщина, элитарщина, аномальщина, слововыражение, покалечение, примаскированный, доследственный, спецлаборатория, промпроизводство, медсправка, педколледж, зоопарикмахер, внеофисный, предшкольный, межкорпоративный, штрафплощадка, видеопоиск, рассинхронизировать, трудносоединимый, Единая Россия, ЕдРо и т. д.

Механизм перекрёстных ссылок, реализованный в словаре, резко увеличивает число синонимов к словам. Поясним на примере. В словарях арго одним из синонимов к слову «оттянувшийся» обычно указывают слово «отдохнувший», но отдельной словарной статьи со словом «отдохнувший» в словарях арго нет. Более того, ни в одном синонимическом словаре среди синонимов к слову «отдохнувший» мы не нашли слова «оттянувшийся». Другие примеры: «схлестнувшийся» — повздоривший, «сдавший» — «выдавший», «сдвинутый» — «ненормальный», «содравший» — «укравший», «убойный» — «замечательный» и т. д. У слов «ассистент», «помощник», «десница» в качестве синонима (толкования) в словарях приводится словосочетание «правая рука». Однако ни в одном словаре словосочетания «правая рука» нет. Электронный (онлайн) словарь-справочник устанавливает все возможные соответствия: синонимами в словаре могут быть не только слова и словосочетания современного общеупотребительного языка, но и устаревшие слова, просторечные, жаргонные, областные, слова профессиональной речи и т. п.

Англо-русские словари разных тематик (научно-технической, программистской, финансовой, медицинской, автомобильной, физической, химической, спортивной, юридической, строительной, биологической и т. д.) существенно использовались для пополнения синонимических связей, а именно: различные переводы английских слов из групп слов этих словарей считались, как правило, синонимически связанными.

5 Исторические параллели

Владимир Иванович Даль закончил петербургский Морской кадетский корпус и несколько лет прослужил офицером на флоте, затем закончил медицинский факультет Дерптского университета и проработал 10 лет армейским хирургом, потом крупным чиновником — действительным статским советником (генерал-майор по воинской табели о рангах), а этнография и лингвистика были для него многие годы любимым увлечением. Прожив 71 год, 53 года он собирал русские слова, выражения, пословицы, поговорки, сказки, песни. Он предложил эти свои материалы Императорской Академии наук для напечатания, но Академия отказала ему в какой-либо поддержке. Пришлось ему самому, «самоучке», заняться упорядочением собранного материала и издать «Толковый словарь

живого великорусского языка» (1863–66 годы), теперь известный каждому образованному русскому человеку. А вот фамилии современников Даля, крупнейших учёных-лексикографов того времени: академиков В. А. Поленова, А. Х. Востокова, М. Е. Лобанова, Я. И. Бередникова, И. С. Кочетова — сейчас и не каждый современный филолог, к сожалению, сможет вспомнить, при всём уважении к трудам этих людей. В словаре Даля имеется около 200 000 слов, и этот словарь называют Русской Библией, а в самой Библии имеется около 13 000 слов.

В. И. Даль писал об отрыве письменного языка от живого русского языка: «Живой народный язык, сберегший в жизненной свежести дух, который придаёт языку стойкость, силу, ясность, целость и красоту, должен послужить источником и сокровищницей для развития образованной русской речи».

Писал Даль В.И. и о сложности выбора порядка в словаре, — алфавитный или корнесловный? При алфавитном способе, как он писал, «я не могу найти слова, которого у меня не хватает; не могу посмотреть сряду самые близкие (сродные) слова, чтобы освоиться с основным значением слов этого корня; не могу отыскать под общим, родовым понятием нужные мне выражения, оглянуть закон и порядок словопроизводства, чтобы осмыслить речь свою… — всё раскинуто врозь; одним словом это не словарь… это список, сборник слов… без связи и смысла, для крайне ограниченного употребления…» «Второй способ, корнесловный, очень труден на деле, потому что знание корней образует уже по себе целую науку и требует изучения всех сродных языков, не исключая и отживших. К тому же этот способ основан на началах шатких, где без натяжки не обойдёшься. Сверх того, при отыскании слов корнесловный порядок предполагает в создателе и читателе одинаковый взгляд и убеждения насчет отнесения слова к тому или иному корню… Поэтому требуется особый, объёмистый указатель и необходимо отыскивать каждое слово дважды, а это докучает и утомляет».

Даль собрал слова по семьям, или гнёздам, родственных по смыслу слов. А в азбучном порядке дал указания, где искать каждое необходимое слово.

В представленном онлайн-словаре, как и в его электронном виде, развитая система запросов позволяет, в отличие от книжных словарей, легко найти требуемое слово и даже помогает осуществлять различные филологические исследования.

Данный словарь-справочник является попыткой продолжения словаря В. И. Даля на современной живой бесцензурной языковой основе (кроме обсценной лексики) с включением слов и словосочетаний из различных отраслей знаний: от Астрономии до Японской живописи. То есть это объединённый словарь русского языка, о необходимости создания которого писали Шишков А. С. и Виноградов В. В. [6].

Словарь-справочник не вызвал никакого отклика у представителей «официальной» лексикографической науки, в печатании статьи по словарю [6] было отказано в 2-х журналах РАН по абсурдным причинам. Это и не удивительно, так как российские «неестественные» науки часто являются существенно клановыми и ценность научной работы определяется её конкретной пользой для какого-либо научного клана; и если автор такой работы не принадлежит каким-либо кланам, то и сама научная работа обычно замалчивается, см. [4].

Электронные версии словаря описаны в статьях [5–7].

6 Идея онлайн-словаря

В марте 2015 года авторы пришли к убеждению, что онлайн-словарь сделать обязательно нужно. К этому времени в Интернете появились онлайн-словари, анонимные авторы которых с нарушением объявленного авторского права «позаимствовали» словарную базу доступного незашифрованного автономного словаря ASIS® версии 6.0

2010 года или ранее, см. выше. Опасение, что словарная база в Интернете может оказаться ещё более уязвимой в отношении несанкционированного копирования, ещё долгое время оставалось.

Но потом стало понятно, что у автора словаря, ежедневно работающего над своим изделием и знающего, где и как искать отсутствовавшие в словаре-справочнике новые слова и словосочетания — несомненное преимущество. К тому же полную версию словаря можно и не размещать в Интернете до поры до времени, не убедившись в надёжности её защиты. Ведь даже версия словаря-справочника десятилетней давности не имеет аналогов в Интернете по составу словаря и количеству синонимических связей.

Предстояло выбрать адекватные технические средства и в полной мере воспользоваться ими.

Важно было с самого начала предусмотреть достаточно быструю, удобную и надёжную технологию обновления словарной базы на сервере, поскольку это может потребоваться в любой момент.

7 Основные принципы реализации

Первое, что требуется разработчику серверных программ, это не ограниченный какими-либо надстройками провайдера доступ к файловой памяти сервера. Предоставление разработчику прав администратора — обычное решение. Настройка FTP для связи компьютера разработчика с выделенным ему серверным ресурсом — оптимальное решение, позволяющее разработчику большую часть дела выполнять «в домашних условиях».

Второе, что необходимо, — это работающий на сервере PHP, желательно не подверженный частым обновлениям, что вынуждает иной раз вносить изменения в надёжно работавшие серверные программы разработчика.

И третье — это согласованная с PHP на сервере реляционная СУБД, в качестве которой в нашей реализации выбрана MySQL.

Принципиальный момент — это не вынуждать Тришина отказываться от хотя и сильно устаревших, но привычных ему по многолетней работе автономных инструментов. При каждой итерации словаря входными данными для разработчика серверных программ служат две большие электронные таблицы — два файла в формате DBF, являющиеся результатом заключительной работы автономных инструментов Тришина.

Существенным недостатком FoxPro, успешно исправленным в MySQL, является невозможность использования текстовых полей переменной длины, которая в процессе развития словаря привела к курьёзам. Так слово русского языка

тетрагидропиранилциклопентилтетрагидропиридопиридиновые

(tetrahydropyranylcyclopentiltetrahydropyridopyridinovyye) потребовало для своей записи в таблице слов 55 букв, а не 50, как было предусмотрено в долго используемой версии автономного словаря.

Ещё более усугубилась ситуация, когда словарную базу пополнили фразеологизмы, толкования и т.п., используемые в качестве синонимов.

Их доля по результатам отдельного исследования не превышает 19,0 %, но значительная их часть не длиннее 50. С учётом этого на автономном словаре пришлось удлинить поле для слов до 200. В СУБД MySQL — это поле переменной длины, и его значение не нужно дополнять «хвостовыми» пробелами.

Таким образом, комплект из FTP клиентской программы на компьютере разработчика, PHP и MySQL на арендуемом сервере — вот и все требуемые инструментальные средства разработчика серверных программ. Автор комплекса

программ воспользовался также своим знанием и профессиональным опытом в применении объектно-ориентированного языка высокого уровня JAVA и его обширной и хорошо описанной библиотекой объектов. Требовалось также знание HTML, CSS и JavaScript, поскольку интерпретатор PHP на сервере передаёт на клиентский компьютер коды именно на этих языках.

Важно, что на сервере реляционная база таблиц слов и синонимических связей должна находиться в строго упорядоченном виде, чтобы ускорить запросы SQL и получать множественные результаты уже в упорядоченном виде. Заниматься упорядочением по запросу любого обратившегося к словарю клиента приводило бы к существенному замедлению отклика на его (их) интерактивные запросы.

8 Создание технологии преобразования из FoxPro в MySQL

Записи в базу, полное обновление таблиц производятся разработчиком только в редких сеансах обновления.

Для обновления таблиц используется комплекс связанных по данным программ, стартующий с чтения и преобразования во внутреннюю текстовую форму таблиц, полученных в виде DBF. На этом же этапе происходит построение таблицы преобразования числовых идентификаторов слов в последовательные номера, начиная с нуля (без пропусков).

Следующий этап — сортировка таблицы слов в алфавитном порядке с использованием особого алгоритма упорядочения. Заглавные (прописные) буквы при упорядочении неотличимы от строчных, буква ‘ё’ при сортировке неотличима от ‘е’. И только в случае равенства слов равной длины она оказывается на своём месте (сразу после буквы ‘е’). Символы пробел и дефис перед сравнением изымаются. Латинские буквы и прочие знаки упорядочиваются после русских букв. Сортировка таблицы в полмиллиона записей выполняется не быстро, при использовании в компьютере Интеловского микропроцессора i7 на её выполнение требуется свыше двух с половиной часов.

Важным этапом является корректировка таблицы синонимических связей по ранее построенной вспомогательной таблице. При этом удаляются строки с несуществующими идентификаторами. Также удаляются при обнаружении строки типа «сам себе синоним». Недостающие обратные связи вставляются. Новые связываемые номера — это просто номера строк-слов, начиная с нуля и подряд с единичным шагом. На эту работу микропроцессору i7 требуется чуть менее двух часов.

Самый продолжительный этап — это сортировка таблицы бинарных синонимических связей (около пяти часов для микропроцессора i7). Оно и понятно: так длина таблицы почти вчетверо больше. Правда, алгоритм сортировки проще. Программа сортирует (упорядочивает) таблицу синонимических связей по возрастанию значений левого числового идентификатора. При равенстве сравниваемых значений выполняется упорядочение по возрастанию значений правого числового идентификатора. Перед сортировкой выполняется перекодировка всех числовых идентификаторов в соответствии с новым порядком, найденным при упорядочении таблицы слов.

Дальнейшие этапы выполняются быстро. В результате получаются программы на PHP, которые затем загружаются в рабочую директорию на сервере и последовательно выполняются в HTTP. Эти программы по частям последовательно загружают новые таблицы вместо удалённых старых таблиц.

Разбиение программ загрузки базы вызвано ограничениями по памяти и быстродействию старых серверов, из-за чего случались тупики при выполнении длинных программ загрузки.

9 Интерфейс пользователя trishin.net

Интерфейс любого посетителя сайта trishin.net с Большим русским словарём-справочником синонимов (близких по смыслу слов), по убеждению авторов, должен быть наиболее простым и понятным. Вместе с тем должна быть предусмотрена достаточно надёжная защита от систематических попыток получить копию словаря в машиночитаемом виде.

Вот почему ответы на запросы к базе с целью усложнения деятельности «копирователей» выводятся на экран не в виде текстов, а в виде ниспадающих списков, которые невозможно скопировать с помощью техники копи-пэйста (copy-paste). Остаётся лишь вручную выбранное слово из списка вводить в текстовое поле для новых запросов к базе.

Таким образом, при первом вызове веб-страницы trishin.net на экран выводится форма с краткими пояснениями, которая имеет пустое поле для заполнения его фрагментом слова для поиска его расширения после выбора мышью изображения кнопки «Найти». Кнопка «Найти» в дальнейшем инициирует выполнение поиска в базе по любому заданному запросу.

Запросы могут отличаться в зависимости от выбора одной из четырёх альтернатив, изображённых небольшими кружками, которые можно отмечать щелчком левой кнопки мыши при наведении курсора внутрь кружка. После щелчка состояние селектора альтернатив соответственно изменяется и сохраняется при всех последующих запросах к словарю вплоть до следующего переключения.

Каждый поиск (кнопка «Найти») начинается с расширения шаблона, находящегося в текстовом поле формы в зависимости от состояния селектора альтернатив. По окончании выполнения запроса к базе, который отрабатывается сервером менее секунды, происходит обновление экрана с сохранёнными состояниями текстового поля и селектора альтернатив, но выводом ниспадающего списка найденных слов. По умолчанию выбранным в списке считается первое слово. В дальнейшем в списке можно отметить курсором (выбрать) другое слово. В форме сообщается число слов в списке, например, «Найдено 250 слов». При этом в ниспадающем списке все 250 слов и представлены.

Для выбранного слова указано число найденных синонимов, а в правом ниспадающем списке все синонимы выбранного слова и находятся. В качестве выделенного синонима так же, как и в левом списке, по умолчанию выбирается первое слово из списка, но этот выбор перед повторным поиском можно с помощью мыши изменить. Справа от выбранного синонима указано число синонимов к этому синониму. Конкретно получить список синонимов синонима можно, введя этот синоним синонима в текстовое поле и повторив поиск.

Если сообщается, что «у этого слова нет синонимов», не удивляйтесь. Значит, так оно и есть.

Так вкратце и устроен интерфейс посетителя сайта trishin.net с «Большим русским словарём-справочником синонимов» (близких по смыслу слов).

10 Заключительные замечания

Этот словарь-справочник практически показывает, что по количеству слов русский язык относится к наиболее развитым языкам мира, а по масштабу и плотности пространства синонимов (а также рифм, интонаций) ему, по мнению авторов, нет равных. Встраивание словаря-справочника в поисковые системы может существенно повысить качество поиска информации.

Литература

[1] Тришин В.Н., Шатров М.В. Система информационной поддержки оценщика ASIS® (Appraiser Support Info System) // Инвестиции в России. 1995. № 5. стр. 35–37.

[2] Тришин В.Н., Шатров М.В. Основные задачи и технические решения, реализованные в компьютерной системе помощи оценщику и аудитору ASIS® // Имущественные отношения в Российской Федерации. 2004. № 11, http://www.trishin.ru/left/publishes/main-task/ .

[3] Тришин В.Н. Это было совсем недавно // в книге «Остров “Информэлектро”: К 60-летию Института», Москва, 2003, под ред. Н.А. Мироновой, стр. 79–80.

[4] Тришин В.Н. Наука и квазинаука. // Вопросы оценки. 2010. №2. стр. 10–28, http://www.trishin.ru/left/publishes/science-and-kvazi/ .

[5] Тришин В.Н. Электронный словарь-справочник синонимов русского языка системы ASIS®. // в книге «Владимир Даль в счастливом доме на Пресне», М.: издательство "Academia", 2010, стр. 158–165.

[6] Тришин В.Н. Мощь русского языка по данным синонимического словаря-справочника системы ASIS®. // Журнал «Вестник УМО. Экономика, Статистика и Информатика», № 6, 2013, стр. 7–13, http://www.audit-it.ru/articles/soft/a119/597862.html

[7] Куянов Ю.В., Тришин В.Н. Количественный анализ Большого словаря-справочника синонимов русского языка. // Журнал «Научное обозрение: гуманитарные исследования», № 9, 2015, стр. 105–111, http://www.audit-it.ru/articles/soft/a118/838916.html.

Приложение. Возможные варианты коммерческой реализации словаря-справочника

1) Переписывание имеющихся электронных версий словаря-справочника в более современные зашифрованные (как для слов, так и для синонимических связей) версии под Windows с улучшенным дизайном, быстрой и легкой установкой словаря, документацией и т.п. для их коммерческого распространения как в интернет-магазинах, так и на дисках.

2) Создание полного аналога словаря для платформы Mac OS.

3) Создание интернетовской версии словаря со всеми поисковыми возможностями версии на Fox Pro. Этот вариант словаря можно будет предложить для использования в интернет-поисковике для рекламы самого поисковика, а также для улучшения качества поиска информации и в качестве подсказчика слов. Для сравнения укажу, что в поисковике Yandex вывешен синонимический словарь Н. Абрамова на 20 тыс. слов, см. http://slovari.yandex.ru/~книги/Словарь%20синонимов/
Также эта онлайн-программа может заинтересовать любую другую компанию с большим числом клиентов в качестве рекламы на их сайте (Евросеть, Связной и т.п.).
4) Создание версий словаря под мобильные устройства для **iOS** и Android (в первую очередь для планшетников) и коммерческое их распространение.

Справка.
Всего в России было продано 7.2 миллиона планшетов общей стоимостью 81.4 миллиарда рублей (к 2014 году), http://www.vladtime.ru/computers/353267-obem-prodazh-planshetov-v-rf-za-2013-god-vyros-na-125.html.
Количество смартфонов, проданных в России в 2013 г., достигло 19,7 млн — на 54,2% больше, чем в 2012 г, см. http://www.vestifinance.ru/articles/41276.

5) Указом Президента РФ В. В. Путина от 09.06.2014 создан президентский Совет по русскому языку. Новый орган, в состав которого вошли как ученые, так и руководители СМИ, должен "защищать" язык и продвигать его за рубежом.
Как говорится в документе, совет является консультативным органом при президенте России. Его задачи — совершенствование госполитики "в области развития, защиты и поддержки русского языка", а также координация работы госорганов, научных и культурных организаций. Кроме того, совет должен заниматься "укреплением позиций русского языка в мире, расширением географии и сфер его применения, поддержкой русскоязычных сообществ за рубежом".

Работа по словарю лежит в русле этого постановления.
В следующей версии электронного словаря версии 8.4 на C# для облегчения изучения работы программы пользователями-иностранцами (осваивающих русский язык) планируется дополнительно переписать интерфейс программы и встроенное руководство пользователя на 7-9 иностранных языках (английском, французском, немецком, испанском и др.), с тем, чтобы в зависимости от установленного при загрузке компьютера языка, этот же язык выбирался для интерфейса и встроенного руководства словаря-справочника.

6) Словарь-справочник удобен и необходим при реализации интеллектуального поиска информации в различных корпоративных системах (структура его словарной базы на языке Fox Pro, в которой кроме слов имеются их основы и коды одного из 19 эталонных слов для указания всех склонений или спряжений слова, проектировалась именно для таких задач).

7) Программа словаря-справочника может послужить хорошим массовым рекламным продуктом для фирмы-изготовителя коммерческих версий словаря путём помещения в программу словаря рекламы других своих наукоемких программных продуктов.

8) Словарь в полностью зашифрованном виде можно разместить в Сети на отдельном сервере или с использованием надёжного "облачного" сервиса с защитой от "выкачки" всего его содержимого. Доступ к словарю мог бы осуществляться через API (Application Programming Interface - Интерфейс Прикладного Программирования), с регистрацией пользователей и продажей ключей доступа на основе той или иной ценовой модели: подписка, квота на количество запросов или что-то ещё. Это откроет возможности использования словаря программами-роботами, которые могут создаваться подписчиками сервиса для различных целей. Одним из вариантов использования может быть интеграция с системой "Антиплагиат", используемой Российской Национальной Библиотекой для проверки оригинальности текстов диссертаций, или подобными ей (для профессиональных журналов, СМИ и др.); другой вариант использования - встраивание в национальные поисковые системы различных стран (Китай, Индия, Бразилия, США, Россия и др.), в которых обработка запросов пользователей должна осуществляться не только на национальном языке, но и на других языках; третий вариант использования - системы автоматического перевода с учётом синонимии (интегрированные с поисковыми системами, как Google Translate - либо с иными информационными продуктами и сервисами); четвёртый вариант использования - поддержка различных приложений для мобильных устройств с сетевыми запросами к словарю по мере необходимости. Доступ к словарю через API - это современный и наиболее гибкий способ публикации, со множеством различных вариантов использования в соответствии с потребностями конкретного пользователя; гибкость в выборе ценовой модели в зависимости от категории и потребностей пользователя также является привлекательной.

Есть и другие бизнес-идеи.

Автор готов передать эксклюзивные права на словарь-справочник компьютерной компании, готовой взяться за изготовление коммерческих версий словаря и их продажу.

Отзывы на словарь с сайта автора.

http://www.trishin.ru/left/dictionary/reviews/