I.Bolshakov, P.Cassidy, A.Gelbukh

CrossLexica - a Dictionary of Collocations and Thesaurus of the General Russian Lexicon *(In Russian).*

// Proceedings of the International Conference on Applied Linguistics "Dialogue-96", Kazan, May 1996, Moscow, Russia.

A dictionary of a new type has been compiled, combining the features of a thesaurus and a dictionary of collocations of the general Russian lexicon. For each word the dictionary presents words semantically related to this word (such as synonyms, antonyms, hypernyms, species, parts, etc.), the derivatives of this word in other parts of speech, and the words that can form collocations with this word. Both idiomatic and frequent free collocations are included. Also presented are inflectional paradigms, government patterns for verbs and nouns, as well as some usage marks. To our knowledge, CrossLexica is the largest available dictionary of Russian collocations: For 77 thousand words, it contains 375 thousand bilateral semantic links and 350 thousand collocations.

A unique feature of this dictionary is its ability to suggest new semantic links and collocations by utilizing the semantic relations between words. This feature effectively adds millions of high-probability links and collocations to the dictionary, useful for those having mostly a passive knowledge of Russian.

A hypertext-style Windows-based program provides convenient access to the dictionary, with Russian-to-English and English-to-Russian translation of individual words and some collocations, as well as English interface, on-line help, and documentation. The dictionary is intended for all people learning or using Russian.

Contacts: gelbukh@micron.msk.ru, cassidy@micra.com.

КроссЛексика - словарь словосочетаний и тезаурус общеупотребительной лексики русского языка

И.А.Большаков, П.Дж.Кассиди, А.Ф.Гельбух

Проф. И.А.Большаковым (Россия) при участии д-ра П.Дж.Кассиди (США) и группы разработчиков составлен словарь нового типа, объединяющий черты тезауруса и словаря словосочетаний общеупотребительной лексики. По слову словарь позволяет определить другие слова, как семантически с ним связанные (не обязательно синонимы), так и сочетающиеся с ним в текстах. Словарь реализован в виде программы для MS Windows и предназ-

начен, в первую очередь, для пользователейДнелингвистов, другими словами, для самого широкого круга пользователей.

Такое назначение словаря определило и его структуру, представляющую, по нашему мнению, удачный компромисс между сложностью известных комбинаторных словарей и простотой использования, необходимой пользователямДнеспециалистам. Многие десятки изученных в лингвистике типов семантических и синтагматических связей были сведены к нескольким основным типам, достаточно близким к тем, которые пользователи "проходили в школе".

Конкретно, в тезаурусе представлены, во-первых, семантические связи - синонимы, антонимы, родовые понятия (гиперонимы), видовые понятия (гипонимы), часть (меронимы), целое (холонимы), семантическа группа (семантические дериваты), во-вторых, синтагматические связи - управляющие существительные, управляющие глаголы, определяющие понятия, определяемые понятия, управляемые слова. Последние разбиты по типам моделей управления.

Например, для слова "вагон" в тезаурусе указаны целое (поезд), части (тамбур, купе), дериваты (вагонный, в вагоне), определяющие понятия (товарный вагон), управляющие глаголы (сесть в вагон, вагон тронулся), управляющие существительные (билет в... вагон), модель управления (вагон до станции...). Для слова "закон" целое - "свод законов", часть - "буква закона", "статья закона", семантическая группа - "юрист", "по закону", и т.д.

Кроме указанных основных функций, в словаре даются пометы об употреблении слов и словосочетаний, такие как "устаревшее", "вульгарное", "идиома" и т.п. Словарь позволяет также просматривать парадигмы слов и задавать исходное слово и даже словосочетание в любой морфологической форме.

Подобный словарь окажется особенно полезным для лиц, изучающих русский язык как иностранный, читающих и составляющих тексты на нем. Для удобства использования иностранцами приводятся переводы слов и идиоматичных словосочетаний на английский язык, кроме того, для омонимичных слов приводятся уточнения, различающие смыслы омонимов, на русском и английском языках. Программа имеет англоязычный интерфейс и документацию и снабжена встроенным англо-русским словарем.

Интересной особенностью словаря является возможность автоматического построения гипотетических семантических и синтагматических связей, что резко увеличивает объем предоставляемой пользователю информации, ценой снижения ее надежности. Так, список связей для слова может быть пополнен связями, которыми обладают его синонимы или гиперонимы. Например, по имеющимся в словаре связям "рвать ягоду" и "крыжовник - гипоним ягоды" может быть автоматически построено и предложено пользователю гипотетическое сочетание "рвать крыжовник", отсутствующее в текущей версии словаря. Естественно, такие гипотетические связи снабжены специальной пометкой, предупреждающей

пользователя о ненадежности данной информации.

Конечно, при этом возникают неправильные сочетания типа "*вызвать медика" (врача) или "*беспозвоночный волк" (животное), и вопрос их отсева - предмет отдельной работы. Если пользователь хотя бы пассивно владеет языком в достаточной степени, чтобы судить о правильности или неправильности таких словосочетаний, данная возможность окажется ему весьма полезной, поскольку число "правильных" гипотетических словосочетаний все же значительно превышает число неправильных. Гипотетические связи представляют собой богатейший, хотя и сырой, материал для дальнейшей работы над словарем, и являются удачным примером совместного использования функций тезауруса и словаря словосочетаний.

Отличительной чертой данного словаря по сравнению с другими является его значительный объем, во много раз превышающий объем известных нам источников подобного типа. Словник словаря составляет 77 тысяч слов, количество семантических функций - более 750 тысяч, синтагматических - около 700 тысяч (по связям в одном направлении; двунаправленных связей примерно вдвое меньше). Для сравнения можно привести отличия данного словаря от известного аналогичного продукта "Русский филолог":

	Русский филолог	КроссЛексика
Объем словника	100 тыс.	77 тыс.
Парадигмы	Зализняк	частичные
Определения	Ожегов	нет
Синонимы/антонимы	Евгеньева	есть
Сем. дериваты	нет	есть
Часть/целое	нет	есть
Род/вид	нет	есть
Модель управления	глаголы	глаголы и существит.
Примеры управления	нет	есть
Словосочетания	нет	350 тыс.
Примеры употребления	нет	в стадии разработки
Перевод на английский	нет	есть

Упоминая в графах таблицы фамилии авторов широко известных печатных изданий, мы не имели целью преуменьшить проведенную авторами "Русского филолога" работу, а желали лишь подчеркнуть, что большая часть информации соответствующих частей этого программного продукта, в принципе, может быть найдена в опубликованных изданиях. В то же время основную часть нашего словаря составляет оригинальный материал, нигде ранее не опубликованный и, насколько нам известно, не повторяющий никакой аналогичный источник.

Как уже отмечалось, словарь предназначен в первую очередь для использования конечным пользователем. Однако его можно рассматривать и как ценный лексикографический ресурс для использования в программах автоматической обработки текста.

Возможности использования тезауруса для задач информационного поиска, извлечения информации из текста и т.п. хорошо известны, однако в прошлом создатели русских тезаурусов ориентировались в основном на научно-техническую лексику, поэтому в настоящее время, когда резко возрос интерес к обработке текстов неспециального содержания, ощущается нехватка машинных тезаурусов общеупотребительной лексики. Этот пробел мы и старались заполнить.

Пополнение машинного тезауруса значительным количеством синтагматических связей и устойчивых словосочетаний придает ему новое качество и новые широкие возможности использования. Одна из сфер применения такого словаря - автоматическая оценка грамматичности и стиля текста путем подсчета в нем числа известных системе словосочетаний и выявления словосочетаний, отсутствующих в словаре. Более серьезное и интересное применение - автоматический синтаксический анализ, опирающийся на поиск в тексте известных словосочетаний.

Кроме того, мы рассматриваем накопленный материал как богатый источник данных для дальнейшей более тонкой классификации, составления словарей определенного типа, например, идиоматического словаря, и другой лексикографической работы.

Словарь существует и может использоваться, во-первых, в виде ориентированной на конечного пользователя программы под MS Windows с развитым пользовательским интерфейсом, во-вторых, в виде набора программ и программных интерфейсов на языке Си++ для использования в прикладных программах, в-третьих, в виде исходных текстов собственно словарных массивов в машинно-читаемой форме и алгоритмов формирования дополнительных связей. Синтаксис исходных текстов словаря достаточно прост и естественен, что позволяет использовать эти текстовые массивы и отдельно от программ и алгоритмов. Общий объем словаря составляет около 8 мегабайт.

14-фев-95