I.Bolshakov, P.Cassidy, A.Gelbukh

Parallel English and Russian Hierarchical Thesauri with Semantic Links, Based on an Enriched Roget's Thesaurus *(In Russian)*.

// Proceedings of the International Conference on Applied Linguistics "Dialogue-95", Kazan, May 1995, Moscow, Russia.

A new lexicographic resource has been developed: FACTOTUM, a large hierarchical thesaurus of the general lexicon, in parallel English and Russian versions.

The English thesaurus was initially based on Roget's thesaurus. The vocabulary entries have been arranged in a deep hierarchy which was reorganized and expanded from Roget's to optimize inheritance, and new words have been added. About 15,000 explicit semantic links of over 100 types such as {antonym}, {property}, etc. have been introduced. By the size and hierarchical structure, FACTOTUM is comparable to the existing WordNet system, however the lexicons of the two thesauri substantially differ, especially in the phrases. WordNet has fewer types of semantic marks, but also contains certain functions absent in FACTOTUM.

The Russian thesaurus was obtained mostly by translating the English part; most of the features described above have been preserved. To our knowledge, it is the largest available hierarchical thesaurus of the general Russian lexicon. Although each of the two parts may be used independently, the bilingual parallel thesauri together form a unique resource for learners and translators, as well as for computational linguists and programmers.

Contacts: cassidy*micra.com, gelbukh*gelbukh.com

РУССКИЙ РОЖЕ:
ПАРАЛЛЕЛЬНЫЕ РУССКИЙ И АНГЛИЙСКИЙ ВАРИАНТЫ ИЕРАРХИЧЕСКОГО ТЕЗАУРУСА С СЕМАНТИЧЕСКИМИ СВЯЗЯМИ НА БАЗЕ ПОПОЛНЕННОГО ТЕЗАУРУСА РОЖЕ

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

Исследовательской группой под руководством проф. И.А.Большакова (Россия) и д-ра П.Дж.Кассиди (США) создан новый лексикографический ресурс: FACTOTUM - иерархический тезаурус общеупотребительной лексики на базе известного тезауруса Роже, параллельные русский и английский варианты.

Основной структурной единицей тезауруса является понятие, выраженное группой синонимичных слов или устойчивых словосочетаний. Понятия сгруппированы в словарные статьи и снабжены пометами, эксплицирующими различные отношения между ними, как семантические, так и контекстные. Словарные статьи расположены в порядке семантической иерархии существительных (прочие части

речи даны в словарных статьях соответствующих существительных), однако множественная подчиненность также учитывается.

Основное внимание уделено отношениям синонимии, семантической близости и гипонимии/гиперонимии. В основном эти отношения задаются иерархическим расположением словарных статей, специальной их нумерацией и группировкой понятий. В случае множественной подчиненности используются явные пометы, например: школьник {is a: учащийся} {is a: ребенок}. Аналогичные пометы используются для указания отношений синонимической близости между понятиями, далекими в иерархии.

Словник английской версии тезауруса включает 45 тыс. слов и 30 тыс. устойчивых словосочетаний (все приводимые цифры - приближенные, поскольку работа по пополнению тезауруса продолжается). Наблюдается существенное различие словников тезауруса FACTOTUM и известного английского тезауруса WordNet, версия 4.0:

	Только FACTOTUM	Пересечение	Только WordNet
Всего:	42646	31901	68758
Отдельные слова:	15981	26289	30147
Словосочетания:	26665	5612	38611
Существительные:	21923	19104	51738
Прилагательные:	7227	6613	7325
Наречия:	2458	814	1886
Глаголы:	10420	5370	7809
Другие части речи:	618	0	0

Отношения и лексические функции, не сводимые к синонимии и гиперонимии, даются в виде явных помет, например: typist {uses: typewriter}. В тезаурусе используется ок. 400 различных отношений и лексических функций, из них 200 используются не менее трех раз, и 25 - не менее ста раз. Следующие группы помет представлены наиболее широко (указано количество):

```
4600: {is a} (дополнительные гиперонимы)
2000: {has_property}, {trait_of}
1300: {causes}, {uses}, {function_of}, {performed_by}
1000: {has_part}, {has_conceptual_part}
 800: [of], [in_frame] (контекстные отношения)
 500: {similar_to}, {antonym_of}
 200: {location_of}
 ...
```

Всего в тексте дано ок. 15000 явных помет (в данный перечень не вошли отношения синонимии, семантической близости и иерархические отношения, задаваемые группировкой и специальной нумерацией понятий и словарных статей). По-видимому, количество

различных семантических отношений может быть сокращено за счет их унификации.

Английский вариант тезауруса составлен д-ром П.Дж.Кассиди. Первоначально за основу было взято издание тезауруса Роже 1913 года (поскольку более поздние издания защищены авторским правом). Текст был пополнен значительным количеством современных слов и понятий, структура словарных статей была обновлена и во многих случаях пересмотрена. Однако основной частью работы явилось построение иерархической структуры и добавление эксплицитных семантических связей и лексических функций. Английский вариант тезауруса отражает в основном языковые нормы, принятые в США.

Русский вариант тезауруса получен в основном путем перевода английской версии. Основной единицей перевода являлось не отдельное слово, а синонимическая группа - понятие, переводом которого считалось множество слов и устойчивых словосочетаний русского языка, имеющих соответствующий смысл в одном из своих значений. Семантические отношения между понятиями, как правило, оказывались сохранившимися при переводе, однако, где это было необходимо, учитывалась специфика русского языка.

Таким образом, синонимические группы, семантические связи и иерархическая структура русского и английского вариантов находятся в близком соответствии друг с другом, что делает FACTOTUM уникальным в своем роде двуязычным лексикографическим источником. Стоит заметить, что еще в 1852 г. о необходимости создания подобных многоязычных тезаурусов писал сам П.Роже, говоря о возможности создания универсального языково-независимого тезауруса.

Русский перевод тезауруса выполнен группой лингвистов под общим руководством проф. И.А.Большакова. В процессе работы над переводом накоплен как положительный, так и отрицательный опыт в области составления двуязычного тезауруса. Вскрылся ряд проблем, связанных как собственно с процессом перевода, так и с различием в структуре языков. Так, существует тенденция к различному дроблению понятий в двух языках при их классификации по какому-либо признаку. Далее, ряд явлений, выражаемых лексически в одном из языков, оказывается выражен регулярным словообразованием или даже словоизменением в другом. При классификации по двум признакам могут быть различия в оценке того, какой из признаков является более значимым. Подобные явления могут приводить к различиям в иерархии понятий двух языков.

Планируемая область применения тезауруса весьма широка. Английская и русская части тезауруса могут быть использованы как совместно, так и по отдельности в качестве одноязычных словарей, а также применяться непосредственно пользователем-человеком или в составе системы автоматической обработки текста.

При использовании человеком одноязычный тезаурус может применяться для подбора вариантов слов при составлении текста,

как справочное пособие по различным областям знаний, и т.п. Особенно полезен такой тезаурус для изучающих английский (соответственно русский) язык, читающих и пишущих на нем. Иерархическая структура позволяет изучать семантическое поле слова и даже помогает определять смысл незнакомых слов по их синонимам, гиперонимам и семантическим связям.

Однако основной целью создания тезауруса (в первую очередь английского варианта) было использование его в системах автоматической обработки текста. Информация об иерархической структуре понятий позволяет использовать наследование свойств гиперонимов, облегчает информационный поиск и извлечение информации из текста.

Двуязычный вариант тезауруса является ценным источником для изучающих иностранный язык, позволяя изучать семантическое поле иностранного слова с "синхронным" переводом на родной язык. Его удобно использовать для перевода с иностранного языка и в особенности на иностранный язык. Действительно, традиционные двуязычные словари обладают тремя недостатками.

Во-первых, при слове указаны переводы разных его значений, зачастую с неопределенной пометкой "в разных значениях". Во-вторых, переводы синонимичных слов разбросаны по разным частям словаря. В-третьих, такие переводы иногда оказываются слишком привязаны к оттенкам смысла слов. Так, переводя "братья и сестры", пользователь смотрит отдельно "брат - brother" и "сестра - syster", а слово sibling `ребенок тех же родителей' оказывается потерянным. Напротив, при использовании двуязычного тезауруса пользователь в одной словарной статье видит все способы выражения выбранного смысла, и именно варианты перевода, выражающие нужный смысл.

Двуязычный тезаурус может быть использован и при автоматической обработке текста, в частности при автоматическом переводе, реферировании или в задачах выравнивания параллельных текстов (text alignment).

Тезаурус существует в машинно-читаемой форме в виде текстовых файлов. Синтаксис текста определен достаточно строго для автоматической обработки. Тезаурус снабжен составленными А.Гельбухом программами для автоматической проверки, индексирования и конвертирования текста, а также интерактивными средствами поиска слов и просмотра в гипертекстовом режиме словника, иерархической структуры и основного текста. Программы предоставляют также интерфейс со структурой тезауруса на языках Си++/Си для использования тезауруса в прикладных программах.

10-янв-1995