**Мозговой М.В.**
Санкт-Петербургский государственный университет

# Контекстно-ориентированный тезаурус русского языка

*Рекомендовано к публикации профессором Тузовым В.А.*

**Введение.** Большую помощь в создании и редактировании текстов (как художественных, так и технических) могут оказать словари синонимов — *тезаурусы*. Например, среди филологов широкой известностью пользуются словари Е. Александровой [1] и Ю. Апресяна [2]. С развитием компьютерных технологий стали появляться электронные версии тезаурусов. Так, в настоящее время существует и постоянно развивается словарь ASIS [3]. Достаточно развитый тезаурус входит в поставку текстового процессора MS Word.

Всем существующим электронным словарям свойственен один и тот же недостаток: отсутствие учета контекста слова при подборе его синонимов. Так, если в MS Word выделить слово *коса* во фразе *вдали виднелась песчаная коса*, будут выведены все найденные синонимы слова *коса* для каждого из возможных его трактовок, хотя из контекста ясно, что речь идет об отмели около берега водоема. Невозможность распознать правильное значение многозначного слова, вероятно, не столь критична, поскольку количество разных значений одного и того же слова обычно невелико. Однако список неверно выбранных синонимов пополняют и другие слова, в какой-либо форме совпадающие с анализируемым. Так, в приведенном выше примере в список синонимов попадет и слово *коса* в значении *косоглаза*.

Разрешить эту проблему можно при помощи алгоритмов, определяющих смысл слова в данном контексте (этому направлению посвящены конференции Senseval, см. напр. [4]). Для русского языка можно применить, в частности, семантический анализатор В. Тузова [5]. Кроме модуля определения смысла слова семантический анализатор содержит формальный словарь, который может быть использован в качестве словаря синонимов.

Данная работа посвящена изучению применимости семантического анализатора в задаче построения качественного тезауруса для русского языка.

**Выявление значения слова анализатором.** В процессе анализа предложений входного текста семантический анализатор генерирует два информационных блока. Первый блок содержит скобочное описание каждого предложения, отражающее его структуру. Второй блок с помощью специального формального языка описывает значение каждого отдельного слова документа (с учетом контекста). Так, слово *коса* в контексте *вдали виднелась песчаная коса* будет описано с помощью формулы:

```
КОСА S1>Copul(S1:БЕРЕГ\$122416(Z1,Z2),УЗКИЙ\$12/01406)
```

Другим значениям слова *коса* сопоставлены другие семантические описания.

Этот пример показывает, каким образом можно использовать семантический анализатор в качестве модуля, выявляющего корректное значение слова в данном контексте. Подключив анализатор к любому существующему электронному словарю, можно сделать его контекстно-ориентированным.

**Семантический словарь как словарь синонимов.** Семантически близким словам (т.е. синонимам) в словаре В. Тузова сопоставлены схожие семантические формулы. Поэтому, располагая семантической формулой некоторого слова, можно автоматически извлечь схожие по смыслу слова из семантического словаря. Впервые эта идея была описана в работе [6], однако в ней не рассматривался контекстно-ориентированный поиск и, кроме того, не был предложен общий алгоритм построения запросов для извлечения синонимов произвольного слова.

Задача определения семантической схожести слов в рамках модели В. Тузова сводится к определению близости семантических формул. Хотя в общем случае задача определения эквивалентности выражений неразрешима (по Тьюрингу), в нашей ситуации даже сравнительно простые эвристические алгоритмы приводят к хорошим результатам.

Семантическая формула характеризуется прежде всего двумя основными видами своих составляющих: номерами классов и списком базисных функций. Например, в приведенной выше формуле слова *коса* присутствуют номера 122416 и 12/01406, а также базисная функция *Copul* (слова БЕРЕГ и УЗКИЙ являются всего лишь расшифровками значений номеров классов). Аналогично, слово *косой* в

смысле «косоглазый» описано с помощью базисной функции $Hab$ и классов 124, 12/02051 и 124/4112:

```
КОСОЙ S1>Hab(S1:ЖИВОЙ$124,КОСОЙ$12/02051(ГЛАЗ$124/4112))
```

Таким образом, мы можем составить упрощенное описание слова, перечислив номера классов и базисные функции, входящие в его семантическую формулу (сначала указываются классы, затем — функции):

```
КОСА 122416 12/01406 Copul
КОСОЙ 124 12/02051 124/4112 Hab
```

Теперь близость семантических формул можно оценить, используя расстояние Левенштейна [7]. Эксперименты показывают, что расстояние, равное нулю (точное совпадение упрощенных формул) почти всегда сигнализирует о найденных синонимах. Расстояние, равное единице, в ряде случаев также обеспечивает хорошее смысловое приближение. Расстояние, превосходящее или равное двум, обычно уже свидетельствует об отсутствии синонимичности (см. табл. 1).

**Таблица 1.** Примеры найденных синонимов

| Слово | Найденные синонимы |
|---|---|
| коса (песчаная) | 0: —<br>1: крутобережье крутояр рабатка |
| косой (косоглазый) | 0: косоглазый<br>1: кривоглазый |
| девушка | 0: дивчина<br>1: кралечка русалочка снегурочка (и др.) |
| ключ (источник) | 0: родничок фонтанчик<br>1: прудик ручеек речка (и др.) |

нтересно отметить, что «близость» в терминах семантического словаря далеко не всегда совпадает со смысловой близостью, отраженной в обычных словарях синонимов. Например, слову *косой* (в смысле *косоглазый*) достаточно близки слова *криворогий* и *кареглазый*. Семантическая формула слова *косой* может быть расшифрована как «некто, имеющий косые глаза». Эта формула достаточно близка к схожим формулам «некто, имеющий карие глаза» и «некто, имеющий кривые рога».

**Практическая реализация тезауруса.** Расстояние Левенштейна интересно для экспериментирования, однако для практических целей его постоянное вычисление может оказаться слишком ресурсоемким. Поскольку большинство синонимов извлекаются при нулевом расстоянии между упрощенными семантическими формулами, в первой версии тезауруса можно ограничиться точным равенством формул.

Для описываемой программы семантический словарь был преобразован в таблицу формата MS Access. Первый столбец таблицы содержит некоторое слово, а второй — упрощенную семантическую формулу из словаря. Имея упрощенную формулу анализируемого слова (получаемую с помощью простого преобразования полной формулы, предоставляемой семантическим анализатором), с помощью простого SQL запроса

```
select * from MainTable where descr='ФОРМУЛА'
```

можно получить список его синонимов.

**Заключение.** Качество существующих словарей синонимов может быть повышено с помощью модулей, определяющих смысл слова в заданном контексте. Для русскоязычных текстов в роли такого модуля может выступать семантический анализатор В. Тузова. В своей работе семантический анализатор использует формализованный семантический словарь, который может быть использован также в качестве словаря синонимов, легко подключаемого к любым программным продуктам. Семантический словарь, однако, не разрабатывался специально для этой цели, поэтому более качественного результата можно добиться при помощи связки семантического анализатора с каким-либо известным существующим словарем синонимов. С другой стороны, информация из словаря синонимов также может быть добавлена в семантический словарь.

## Литература


1. Александрова З.Е. Словарь синонимов русского языка. М.: Русский язык, 2001. 495 с.

2. Новый объяснительный словарь синонимов русского языка / Под ред. Ю.Д. Апресяна. М.: Языки славянской культуры, 2003. 624 с.

3. Голубицкий С. Побочный продукт // Компьютерра. 2004. № 26–27.

4. Edmonds Ph., Kilgarriff A. (Eds.) Journal of Natural Language Engineering (Special Issue Based On Senseval-2), 2003. V. 9(1).

5. Тузов В.А. Компьютерная семантика русского языка. СПб.: Изд-во СПбГУ, 2004. 400с.

6. Коробейникова О.В., Порошин В.А. Использование компьютерных словарей русского языка — поиск синонимов посредством SQL-запросов // Процессы управления и устойчивость: Труды 34-й научной конференции аспирантов и студентов / Под ред. Н.В. Смирнова, В.Н. Старкова. – СПб.: Изд-во СПбГУ, 2003. С. 379–384.

7. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР. 1965. Т. 163, Вып. 4. С. 845–848.