119270, Москва, Лужнецкая наб., д. 6,
стр.1, офис 214, ООО «ЭР СИ О»
Тел./факс: (495) 287-98-87
E-mail: info@rco.ru
http://www.rco.ru

Руководство администратора RCO Ling – пакет настройки семантических словарей

Версия 3.0
(Microsoft Windows)

Москва, 2007

В содержание данного документа могут быть внесены изменения без предварительного уведомления. Названия организаций, имена и даты, используемые в качестве примеров, являются вымышленными, если не оговорено обратное.

Содержание

RCO Ling – пакет настройки семантических словарей

Обзор

В этом руководстве содержатся основные сведения по настройке семантических словарей, которые используются в продуктах линейки **RCO** при анализе текста.

Целевой бинарный словарь **LingDct.dat**, используемый в программах **RCO** для высокоэффективной работы, собирается из исходных текстовых файлов словаря при помощи программ, входящих в настоящий пакет. Специфика предметной области, в которой работают приложения **RCO**, может потребовать расширения словарей или изменения принятой системы семантической классификации. Для этих целей и предназначен данный пакет.

Для выполнения установки пакета **RCO Ling** необходимо располагать компьютером с установленной операционной системой **Microsoft Windows NT версии 4.0** и выше.

Введение

Настоящие словари используются при анализе русского текста в продуктах **RCO**. Система словарей описывает различные типы семантических отношений между понятиями текста и способы их синтаксической реализации в тексте, а также позволяет отождествить близкие по смыслу слова-синонимы.

В состав комплекта словарей для семантического анализа входят:

- словарь моделей управления объемом более 15 тысяч слов;
- тезаурус-классификатор объемом более 70 тысяч слов;
- словари общеупотребительной лексики объемом более 10 тысяч слов.

Словарь моделей управления

Общий словарь моделей управления **RCO** используется при синтаксическом анализе текста и включает в себя свыше пятнадцати тысяч наиболее употребительных предикатов русского языка (глаголов, отглагольных существительных и прилагательных, предикативов).

Ниже приведен фрагмент словаря, который описывает ситуацию «*аренда*», семантические роли ее участников (объект, субъект, источник и т.п.) и способы их грамматического выражения в тексте.

```
арендовать noun  В Объект // предмет аренды
арендовать noun  И Субъект // арендатор
аренда           noun  Р Объект // предмета аренды
аренда           noun  Т     Субъект // арендатором
аренда,арендовать noun за В Цена // за цену
аренда,арендовать noun на В Время // на срок
аренда,арендовать noun у Р Источник // у арендодателя
аренда,арендовать noun для Р Назначение // для торговли
```

В строке словаря после каждой группы слов-предикатов с указанием типа управления задается предлог или союз (если таковой имеется) и падеж, в котором должно стоять в тексте подчиненное предикату слово (если это существительное или прилагательное). Далее указываются семантическая роль подчиненного слова и возможные комментарии.

Тип управления обозначает класс зависимых слов – существительное, прилагательное, глагол в инфинитной или в личной форме. В приведенном примере все предикаты управляют существительными (именными группами).

Дополнительно в модели управления для повышения точности синтаксического анализа текста можно перечислить семантические категории, которым должны удовлетворять слова-аргументы предиката. Например, в модели управления

```
арендовать noun И Субъект CAT: name;men name;office animated
```

обозначения *"*`name;men`*"*, *"*`name;office`*"* и *"*`animated`*"* указывают, что в качестве субъекта глагола «арендовать» в именительном падеже могут выступать слова категорий: «*имя собственное*» вместе с категорией «*человек*», «*имя собственное*» вместе с категорией «*организация*» или категории «*одушевленное*». Это соответствует любым наименованиям физических и юридических лиц и всем одушевленным существительным.

Специфика предметной области, в которой работают приложения информационного поиска, может потребовать расширения словаря и изменения классификации типов семантических ролей. Так, может возникнуть необходимость замены общих названий ролей «*субъект*», «*объект*» и «*источник*», выделяемых в словаре для широкого класса действий, на более частные, характеризующие ситуацию аренды: «*арендатор*», «*предмет аренды»* и «*арендодатель*».

Тезаурус-классификатор

Тезаурус русского языка используется для отождествления близких по смыслу единиц текста и их семантической классификации в ходе анализа документов.

Тезаурус **RCO** позволяет описать два типа семантических отношений между словами: синонимические и гипонимические (общее-частное).

Тезаурус представляет одноуровневую иерархию, в которой слова (или целые синонимические ряды) могут объединяться под обобщающими понятиями – гиперонимы. Все слова, имеющие одно обобщающее понятие, носят название гипонимов.

Ниже приведен пример фрагмента тезауруса, описывающий семантическое поле, связанное с понятием воровства. Как видно, три синонимических ряда, соответствующие понятиям «*воровать*», «*воровство*» и «*вор*», объединены под обобщающим понятием-гиперонимом «*воровство*». При этом все члены синонимических рядов «*воровство*» и «*воровать*» будут отнесены к семантическим категориям, обозначенным как *act* и *pres* – действие и передача материального объекта.

```
Verb воровать, красть, тырить,... воровство CAT: act;pres
Noun воровство,кража,татьба CAT: act;pres
Noun вор,жулье,жулик,ворюга,воровка,воришка воровство
```

Синонимы и гиперонимы, указанные в тезаурусе, используются для отождествления близких по смыслу единиц текста в ходе синтактико-семантического анализа и приведения их к унифицированному виду при синтезе текста. Например, синтагмы «*украдена коварными жуликами*» и «*коварная воровка крадет*» посредством синонимических замен будут преобразованы к виду «*коварный вор ворует*», а затем, путем приведения к обобщающему гипероними, примут форму «*воровство коварным вором*».

В состав тезауруса общей лексики русского языка, включенного в **RCO**, вошло около 75 тысяч слов, объединенных в 22 тысячи гипонимических рядов (22 тысячи гиперонимов), в том числе 17 тысяч синонимических рядов, охватывающих 45 тысяч слов.

Словари общеупотребительной лексики

Дополнительно в состав комплекта семантических словарей для анализа текста входят словари с удаляемыми и общеупотребительными словами русского языка. Данные словари используются в качестве фильтров при отождествлении близких по смыслу единиц текста и при выделении значимых элементов текста в ходе синтактико-семантического анализа.

Удаляемые слова

Слова данного типа, в зависимости от части речи, к которой они относятся, обрабатываются при анализе текста следующим образом:

- прилагательные удаляются из всех конструкций текста, позволяя отождествить близкие по смыслу конструкции. Например, именная группа «*покупка всех аналогичных товаров*» будет преобразована в «*покупка товара*», если прилагательные «*весь*» и «*аналогичный*» указаны в числе удаляемых слов;
- все текстовые конструкции, которые содержат существительные и глаголы, указанные в числе удаляемых слов, не выделяются в качестве значимых элементов текста. Например, если слово «*отклонить*» входит в число удаляемых, то из текста «*парламент отклонил проект закона ...*» не будут выделены элементы «*отклонение*», «*отклонение проекта закона*» и «*отклонение парламентом*», однако выделятся прочие элементы: «*парламент*», «*проект закона*», «*проект*», «*закон*».

В стандартные словари включено более 3-х тысяч удаляемых слов.

Общеупотребительные слова

Слова данного типа имеют слишком широкое значение, чтобы характеризовать тематику текста в отдельности, без сочетания с другими словами. Например, слова «*использование*» и «*делать*» сами по себе ничего не говорят о содержании текста, но могут оказаться информативными в сочетаниях «*использование химического оружия*», «*делать деньги*».

Все текстовые конструкции, содержащие данные слова, выделяются в качестве значимых элементов, только когда они содержат еще хотя бы одно не общеупотребительное и не удаляемое слово. Например, если слова «*концепция*» и «*развитие*» входят в число общеупотребительных слов, то из фрагмента текста «*концепция развития сельского хозяйства*» не выделятся элементы «*концепция*», «*развитие*», «*концепция развития*», но выделятся «*концепция развития сельского хозяйства*» и «*развитие сельского хозяйства*», так как эти элементы содержат не общеупотребительные слова. И конечно, выделятся «*сельское хозяйство*», «*хозяйство*» и «*село*».

В стандартные словари включено более 7-ми тысяч общеупотребительных слов.

Установка пакета RCO Ling

Для установки пакета просто скопируйте все файлы дистрибутива в папку **RCO Ling** или любую иную папку на вашем компьютере.

В состав пакета входят следующие файлы:

lingdic.ini	файл с перечнем исходных текстовых словарей, используемых для сборки бинарного словаря **LingDct.dat**
***.txt**	набор исходных текстовых файлов словарей, перечисленных в файле **lingdic.ini**
linglexclass.txt, lingroles.txt, lingconnectors.txt	глоссарии используемых в словарях обозначений
LingDct.dat	бинарный файл семантического словаря, создаваемый из текстовых файлов при помощи приложения-компилятора **LingDicCompiler.exe**
LingDicCompiler.exe	компилятор бинарного словаря – консольное приложение
RCOLing.pdf	файл настоящего руководства

Работа с пакетом RCO Ling

Стандартные семантические словари, входящие в состав лингвистического обеспечения **RCO**, хранятся в текстовом виде в файлах, которые перечислены в файле **lingdic.ini**. Все эти файлы находятся в базовой папке установки **RCO Ling**. Утилита **LingDicCompiler.exe**, запускаемая из этой же папки, собирает из текстовых файлов бинарный файл **LingDct.dat**, который используется продуктами **RCO** при анализе текста.

Для настройки словарей следует ввести или удалить требуемые слова в текстовых файлах при помощи текстового редактора, после чего запустить **LingDicCompiler.exe**. В ходе сборки выдается серия сообщений (файл **errors.txt**), которые позволяют ликвидировать конфликты в словарях, возникающие при связывании слов между собой, а также при стыковке с модулем морфологического анализа.

Пополнение словарей новой лексикой может потребовать сопутствующего расширения словаря морфологического модуля. В этом случае следует обратиться к документации «***RCO Morph – пакет настройки морфологического словаря***».

Подключение словарей: файл lingdic.ini

Файл **lingdic.ini** содержит список файлов словарей, участвующих в сборке, с указанием типа словаря. При подключении новых файлов их имена и тип следует добавить в список.

Формат файла

Каждая строка файла имеет вид:

```
(deleted | common | class | control) : <имя файла>
```

Ключевые слова

Данные слова указывают тип словаря:

- *deleted* – удаляемые;
- *common* – общеупотребительные;
- *class* – тезаурус-классификатор;
- *control* – модели управления.

Комментарии

Знак “//” обозначает начало комментария, который продолжается до конца строки.

Пример файла lingdic.ini

```
deleted: deleted_adj.txt   // удаляемые прилагательные
deleted: deleted_misc.txt  // разные удаляемые слова
common : common_noun.txt   // общеупотребительные существительные
class: lexclass_adj.txt    // тезаурус прилагательных
class: lexclass_noun.txt   // тезаурус существительных
control : control1.txt     // модели управления
```

Файлы-глоссарии: linglexclass.txt, lingroles.txt, lingconnectors.txt

Файлы-глоссарии позволяют задать списки обозначений, используемых для проверки словарной информации на отсутствие орфографических ошибок. При наличии в словарях неизвестных обозначений программа сборки сообщит об ошибках, описанных в разделе «Сборка словарей».

linglexclass.txt – содержит перечень наименований семантических категорий тезауруса и словаря моделей управления. Разделителями служат символы “;”, пробела и табуляции. Написание наименований чувствительно к регистру.

lingroles.txt – содержит перечень наименований семантических ролей словаря моделей управления. Разделителями служат символы “;”, пробела и табуляции. Написание наименований чувствительно к регистру.

lingconnectors.txt – содержит перечень предлогов и союзов, используемых в словаре моделей управления. Каждая строка может содержать только один предлог или союз. Их написание не чувствительно к регистру, расстановке пробельных символов и знаков препинания.

Во всех словарях знак “//” обозначает начало комментария, который продолжается до конца строки.

Настройка словаря моделей управления

В списке словарей (файл **lingdic.ini**) словари данного типа указываются с ключевым словом “control”.

Формат файла

Каждая строка файла задает модель управления для набора слов и имеет вид (в нотации БНФ):

```
<слово>(,<слово>)* <тип управления> (<коннектор>)? <падеж>? (CAT:
   (<CategoriesList>)+ )?                          ,
```

где

```
CategoriesList ::= (<категория>(;<категория>)*)+
```

Здесь символ “*” означает ноль или более повторений выражения в круглых скобках, “+” – одно или более повторений, а “?” – ноль или одно повторение.

Коннектор представляет собой предлог или союз, в том числе из нескольких слов, разделенных пробелами и знаками препинания.

Выражение **`CategoriesList`** определяет перечень семантических категорий, к которым должно относиться слово в грамматической форме, заданной моделью управления. В качестве категории может быть также указано любое слово русского языка. Списков категорий или отдельных слов может быть несколько, разделенных символом табуляции. В этом случае модели управления удовлетворяют лишь слова, относящиеся ко всем категориям, приведенным в любом из данных перечней. Если список категорий пуст, модели управления отвечает любое слово в заданной грамматической форме.

Для разделения элементов строки используется только символ табуляции.

Все слова должны быть указаны в нормальной форме. Управление прямыми и возвратными формами глаголов должно задаваться отдельно, при этом во втором случае – в возвратной форме инфинитива.

Все используемые предлоги и союзы должны быть указаны в файле-глоссарии **lingconnectors.txt**. Все используемые названия семантических категорий должны быть указаны в файле-глоссарии **lingroles.txt**. В противном случае приложение сборки словаря выдает сообщение об ошибке.

> **Замечание.** При анализе семантические категории слов извлекаются из тезауруса-классификатора или модуля выделения объектов в тексте (см. «***RCO Pattern Extractor – руководство администратора***»). Если категория слова не определена, считается, что оно удовлетворяет всем ограничениям в моделях управления, кроме представленных конкретными словами.

Написание типов управления, падежей и названий семантических категорий чувствительно к регистру. Написание слов, предлогов и союзов нечувствительно к регистру, за исключением аббревиатур и имен собственных, которые должны писаться с большой буквы.

Ключевые слова

CAT – начало списка семантических ограничений.

noun, verb, inf, adj – тип управления (см. раздел «Типы управления»).

И, Р, Д, В, Т, П – обозначения соответствующих падежей (см. раздел «Падежи»).

Комментарии

Комментарий, продолжающийся до конца строки, начинается с метки “//”.

Пример файла словаря моделей управления

```
// подписал -> руководитель, ООО “Ромашка” <- подписало
подписать noun И субъект CAT: name;men     name;offic
// агитировать -> (с помощью) листовок
АГИТИРОВАТЬ,САГИТИРОВАТЬ noun с помощью Р способ
ГОТОВНОСТЬ inf   // готовность -> исполнить
Казаться,показаться   adj Т признак // казаться -> сильным
видеть,увидеть verb   как // увидеть -> (как) произошло
```

Настройка тезауруса-классификатора

В списке словарей (файл **lingdic.ini**) словари данного типа указываются с ключевым словом “*class*”.

Формат файла

Каждая строка файла имеет вид (в нотации БНФ):

```
<часть речи> <слово>(,<синоним>)* (<базовый термин>)?
(CAT: CategoriesList)?                          ,
```

где

```
CategoriesList ::= <категория>(;<категория>)*
```

Здесь символ “*” означает ноль или более повторений выражения в круглых скобках, “+” – одно или более повторений, а “?” – ноль или одно повторение.

Все слова должны быть в нормальной форме, указанная часть речи – соответствовать части речи слова из морфологического словаря, синонимы – принадлежать к одной части речи. Используемые коды частей речи приведены в Приложении 1. В качестве обобщающего слова (гиперонима) допустимо использование только существительных и эквивалентных им частей речи, имеющих коды **`Noun`**, **`Name`**, **`GeoName`**, **`Abbr`**, **`Appelation`**.

При анализе текста все слова синонимического ряда вначале приводятся к главному синониму – слову, которое указано первым в строке синонимов. Далее в некоторых синтаксических конструкциях главный синоним может быть приведен к базовому термину.

Замечание. Помимо тезауруса-классификатора, замена синонимов может также производиться в модуле выделения объектов в тексте (см. документацию «***RCO Pattern Extractor – руководство администратора***»).

Названия используемых семантических категорий необходимо указать в файле-глоссарии **linglexclass.txt**, иначе приложение сборки словаря сообщит об ошибке.

Написание кодов частей речи и названий семантических категорий чувствительно к регистру. Написание слов нечувствительно к регистру, кроме аббревиатур и имен собственных, пишущихся с большой буквы.

Замечание. Помимо тезауруса-классификатора, при анализе текста семантические категории некоторых слов могут проставляться в модуле выделения объектов в тексте (см. «***RCO Pattern Extractor – руководство администратора***»). В этом случае категория из тезауруса игнорируется.

Ключевые слова

CAT – начало списка категорий.

Комментарии

Комментарий, продолжающийся до конца строки, начинается с метки "`//`".

Пример файла тезауруса-классификатора

```
Adjective выносливый,двужильный выносливость // нет категории Noun
   воровство,кража CAT: f;act;pres // нет обобщающего
Verb воровать,красть,украсть воровство CAT: f;act;pres
Verb вкладывать,вложить,влагать вклад CAT: f;act;pres
Verb воспламенить воспламенение CAT: f;act;emot f;act;phys
Verb воспламенить CAT: f;act;emot f;act;phys
Noun воспламенение CAT: f;act;emot f;act;phys
Verb воспламенять,воспламенить воспламенение
```

Замечание. Допустимо указание одного слова в нескольких строках файла, что иллюстрирует приведенный пример. То есть можно в одной строке указать синонимы, в другой – базовый термин, а в третьей – категории. Общее требование, которое при этом не должно нарушаться, состоит в согласованности и однозначности информации, указанной в различных строках словаря, – слова, к которым приводятся другие слова, сами не должны приводиться к другим словам. Так, транзитивность вида «*украсть,красть*» и «*красть,воровать*», рефлексивность вида «*кража воровство*», «*воровство кража*» в тезаурусе не допускается! В случае подобных нарушений приложение сборки словаря выдает предупреждение.

Настройка словарей общеупотребительной лексики

Словари общеупотребительной лексики содержат удаляемые и общеупотребительные слова русского языка, специфика обработки которых описана в разделе «Введение. Словари общеупотребительной лексики».

В списке словарей (файл **lingdic.ini**) словари удаляемых слов указываются с ключевым словом "*deleted*", а словари общеупотребительных слов – с ключевым словом "*common*".

Формат записи словарей

Каждая строка файла удаляемых или общеупотребительных слов имеет вид:

```
<слово>    (SP: <часть речи>)?
```

Указание части речи не обязательно, но желательно, так как помогает разрешить омонимию в случае ее наличия (например, «*стать*» – нормальная форма как глагола, так и существительного).

Комментарии

Комментарий, продолжающийся до конца строки, начинается с метки "//".

Пример файла

```
стать       SP: Verb   // удаляемый глагол
стать       SP: Noun   // общеупотребительное существительное
АНАЛОГИЧНЫЙ                  // удаляемое слово
Четверть                // удаляемое слово
```

Сборка словарей – приложение LingDicCompiler

Сборки новой версии бинарного словаря **lingdct.dat**, который использует **RCO**, производится из консоли запуском приложения **LingDicCompiler.exe**, расположенного в базовой папке установки пакета **RCO Ling.**

Сообщения об ошибках и предупреждения в ходе сборки помещаются в файл **errors.txt**. Сообщения могут быть вызваны следующими причинами:

- невозможностью открыть один из файлов-глоссариев **linglexclass.txt**, **lingroles.txt**, **lingconnectors.txt** (указанный файл отсутствует в папке запуска **LingDicCompiler.exe**);
- невозможностью открыть файл, указанный в списке **lingdic.ini** (неверно задано имя файла или файл отсутствует);
- приведенная в словаре семантическая категория, семантическая роль или коннектор связи отсутствует в соответствующем файле-глоссарии **linglexclass.txt**, **lingroles.txt**, **lingconnectors.txt**;
- неверно указана часть речи, тип управления или падеж (список используемых кодов приведен в Приложении 1);
- указанное слово отсутствует в словаре морфологического анализа, задано не в нормальной форме или указана неверная часть речи. Возможно, стоит поставить слово в нормальную форму или изменить часть речи либо слово, указанное в качестве обобщающего, не является существительным. Если же слово действительно отсутствует в морфологическом словаре, его необходимо добавить в **MorphDct.dat**, обратившись к документации «***RCO Morph – пакет настройки словаря морфологического анализа***», после чего собрать словари семантического анализа заново;

- слово имеет другой синоним, другой базовый термин, или базовый термин имеет другой синоним. Нарушены требования однозначности при задании отношений в тезаурусе (см. раздел Тезаурус-классификатор).

Причина прочих ошибок – нарушение формата записи файлов.

Приложение 1. Используемые обозначения

Ниже перечислены обозначения, используемые в словарях.

Коды частей речи

В таблице приведены допустимые коды частей речи, используемые в тезаурусе и словарях общеупотребительной лексики.

Код	Часть речи	Пример
Unknown	Неизвестна	-
Introductory	Вводное слово	скорее, однако
Interjection	Междометье	ах
Predicative	Предикатив	трудно, нельзя, был
Preposition	Предлог	для, посредством
Conjunction	Союз	что, иначе
Particle	Частица	же, уж
Adverb	Наречие	скорее, тепло
Comparative	Обособленная сравнительная степень	скорее
Pronoun	Местоимение	что, ты
Numeric	Числительное	несколько, двадцать, пятый
Abbr	Аббревиатура	МГУ, РФ
Name	Имя собственное	Вася, Иванов, Петрович
GeoName	Географическое название	Крым, Москва
Verb	Глагол	длиться, быть
Adjective	Прилагательное	быстрый, свой
Noun	Существительное	использование, уж
Appelation	Наименование	Газпром, Гарант-Парк

Типы управления

В таблице приведены допустимые коды типов управления, используемые в словаре моделей управления.

Код	Тип управления	Пример
noun	управление существительным (с предлогом или без него)	*арендует -> машину, аренда -> (у) автобазы, человеку <- нельзя*
inf	управление глаголом в инфинитиве	*учил -> стрелять, попытка -> найти, можно -> играть, готовый -> ударить*
verb	управление глаголом в личной форме в придаточном предложении (с союзом)	*слышал -> (что) произошло, увидел -> (как) случилось*
adj	управление прилагательным	*казаться -> веселым, была -> свежая*

Падежи

В словаре моделей управления допустимы следующие обозначения падежей:

И – именительный, *Р* – родительный, *Д* – дательный, *В* – винительный, *Т* – творительный, *П* – предложный.