Автоматическое разрешение семантической неоднозначности в Национальном корпусе русского языка[1]

Б.П. Кобрицов, О.Н. Ляшевская
Отдел лингвистических исследований
ВИНИТИ РАН, Москва
neuralman@yandex.ru, olesar@mail.ru

В рамках программы по созданию Национального корпуса русского языка осуществляется пилотный проект комплексной морфологической и таксономической разметки текстов. Таксономическая разметка предоставляет исследователям новый инструмент для решения различных лингвистических задач. Пользователь может осуществлять поиск примеров не только по отдельным лексемам русского языка, но и по группам слов, принадлежащих к общему таксономическому классу (семантическому полю), например, движение, эмоции, размер, время, вещество и т. п.

Данная работа посвящена разработке лингвистического обеспечения для корпуса современных русских текстов середины XX – начала XXI века (www.ruscorpora.ru). Корпус, разрабатываемый в Российской академии наук под руководством А.М. Молдована и В.А. Плунгяна, содержит сбалансированный набор текстов различных жанров: художественную прозу и драматургию, публицистику, официально-деловые тексты, объявления, записки и т. п. В настоящее время корпус насчитывает порядка 25 млн. словоупотреблений, в перспективе ставится задачи довести это число до 100 млн.

На нынешнем этапе проекта входящие в корпус тексты снабжены метатекстовой аннотацией (информация о жанре текста, его авторе, времени создания и т. п.), морфологической и акцентной разметкой (информация о части речи, грамматических признаках и месте ударения, приписываемая каждому слову); см. подробнее [Сичинава 2002, Плунгян 2004].

Поскольку корпус адресован, в том числе, для лингвистов, которым необходим материал исследования свойств естественного языка, предлагается ввести новый уровень аннотации текстов, а именно, аннотацию словообразовательных и семантических характеристик лексем. Деривационный компонент должен обеспечить разметку слов, представляющих наиболее продуктивные словообразовательные классы, например, имен с диминутивными и аугментативными суффиксами, nomina agentis, nomina feminina, приставочных глаголов, возвратных глаголов и т. д. Семантический компонент, на пилотной стадии проекта, обеспечивает разметку слов по принадлежности к достаточно крупным и традиционно выделяемым в лингвистических исследованиях лексическим классам, таким как "движение", "восприятие", "эмоции", "время", "цвет", "размер", "лицо", "вещество" и др. Совокупность информации о принадлежности слов к

[1] Авторы пользуются случаем высказать благодарность нашим коллегам: руководителю проекта Е.В. Рахилиной, Е.В. Падучевой, В.А. Плунгяну, А.Е. Полякову и С.А. Шарову, принимавшим участие в разработке принципов таксономической аннотации, и особенно Е.Ю. Калининой, Г.И. Кустовой, Д.В. Сичинаве, С.Ю. Толдовой, М.В. Филипенко, О.Ю. Шеманаевой, проделавшим огромную работу по разметке словаря. Программы автоматической разметки текстов на основе словарных данных и комбинированного поиска по морфологическим и семантическим признакам созданы Б.П. Кобрицовым и Д.В. Панкратовым.

словообразовательным и семантическим классам называется в нашем проекте таксономической разметкой текста.

В русских корпусах, насколько нам известно, задачи подобного рода ставятся, в ограниченном объеме, разработчиками Аннотированного корпуса русских текстов Хельсинкского университета ХАНКО (www.slav.helsinki.fi/hanko/; [Копотев, Мустайоки 2003]) и Компьютерного корпуса текстов русских газет (www.philol.msu.ru/~lex/corpus/).

Таксономическая аннотация может быть полезна пользователям корпуса в двух отношениях:

1. для сортировки результатов поиска по типу контекста;
2. для поиска в текстах слов одного лексического класса или определенных лексико-грамматических конструкций.

1. Сортировка выдачи по контекстам

Пользователю должна быть предоставлена возможность сортировки найденных примеров по некоторому удобному ему параметру. Конечно, такая проблема не стоит, если пользователь ищет употребление какого-нибудь редкого слова, например, глагола *раззявиться*. Можно ожидать, что в корпусе найдется менее десятка примеров. Однако при поиске частотных слов или грамматических форм (например, глагола *идти* или существительных в родительном падеже) количество найденных примеров может достигать нескольких тысяч.

Выдача найденного материала может быть отсортирована:

- по жанрам текстов (в порядке приоритета, заданного пользователем), по времени создания и другим метатекстовым характеристикам;
- по алфавитному порядку левого или правого контекста;
- по частеречным и морфологическим признакам левого или правого контекста (см. рис. 1);
- по семантическим классам, в которые входят слова левого или правого контекста (см. рис. 2).

Аргентина	**идет** русским путем ... Осипов Георгий. М
Игорь Трунов тут же пояснил, что речь	**идет** об одном миллионе долларов. ... Бого
Неужели Соколов не понимает, что речь	**идет** о чем-то неизмеримо большем, чем о
Кредитование реального сектора	**идет** ни шатко ни валко. ... Иогансен Ниль
дтвердил "Известиям" Эдуард Кузьмин, все	**идет** по плану ... Авдеев Сергей. 7 суток --
Россия -- страна, которая	**идет** к открытому обществу и не боится кр
что, во-первых, о моей режиссуре и речи не	**идет**, и, во-вторых, как актер я (как, впроч

Рис. 1. Сортировка по морфологическим характеристикам левого контекста

Судьба ведет человека, но человек	**идет** потому, что хочет, и он волен не хоте
И вот уже ребенок	**идет** от лужи, идет с чужим дядей, по-наш
...	
шум, звенело в ушах и все казалось, эшелон**идет**, идет.. ... Василий Гроссман. Все тече	
он, убитый, все жал на акселератор, и танк	**идет**. ... Василий Гроссман. Жизнь и судьб
...	
Впрочем, речь	**идет** не обо мне.. ... Сергей Довлатов. Наш

Рис. 2. Сортировка по семантическим характеристикам левого контекста

Представляется, что хороший результат в организации материала может дать сочетание двух последних типов параметров.

2. Поиск по лексическому классу

Морфологическая разметка в корпусе позволяет пользователю искать как отдельные лексемы или формы, так и их сочетания, например, генитивные конструкции с отрицанием (рис. 3): ср. *жары не чувствовалось*; *зимы не ожидалось*; *ответа не последовало*.

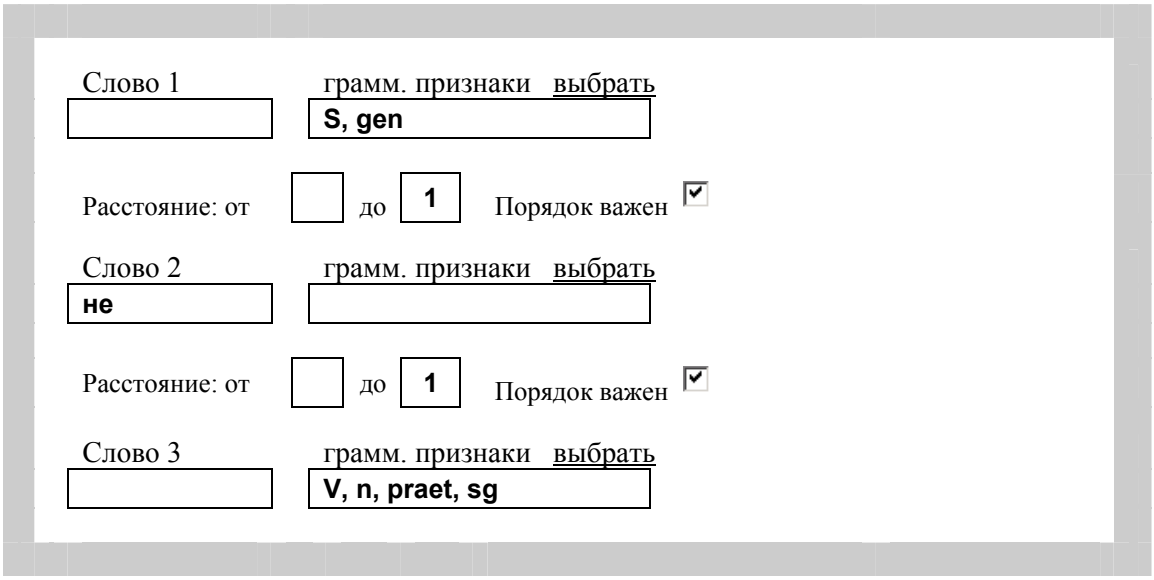

Рис. 3. Поиск по лексико-грамматическим характеристикам

Использование семантических признаков усиливает возможности такого поиска. В частности, можно искать языковой материал, где в данной конструкции выступают глаголы восприятия (рис. 4), ср. *жары не чувствовалось; никакой тревоги не замечалось*.

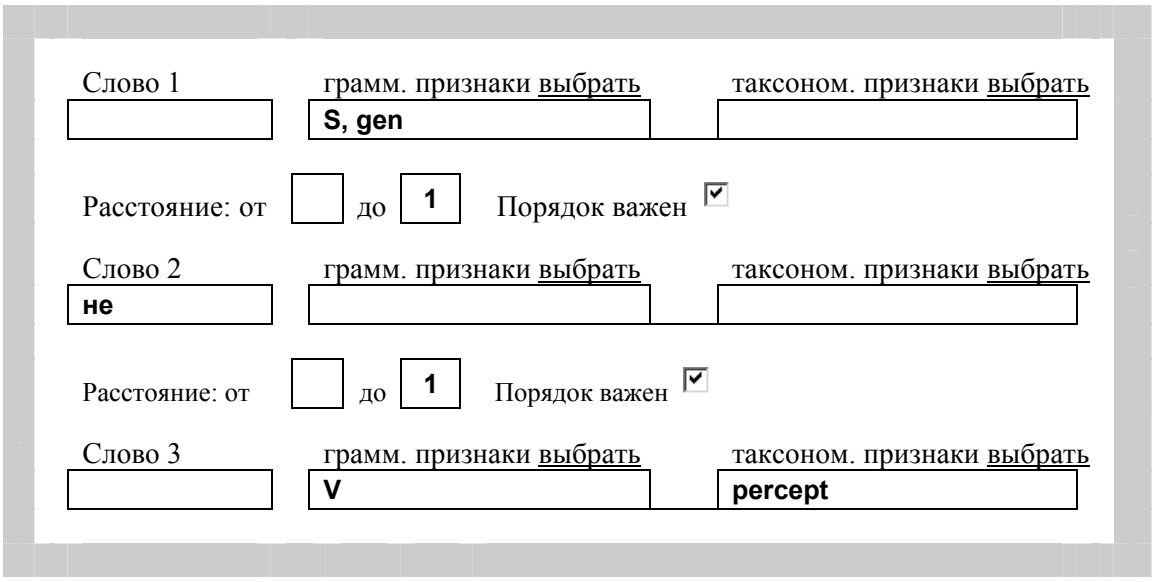

Рис. 4. Поиск по лексико-грамматическим и таксономическим характеристикам

Другой пример: допустим, требуется найти в корпусе употребления параметрической конструкции типа *человек высокого роста*. В случае "простого" лексико-грамматического поиска пользователю понадобилась бы целая серия запросов, в которых он должен был бы последовательно перечислить все слова типа "рост", "вес", "объем", "длина" и т. д. Опция поиска по таксономии позволяет добиться аналогичного результата с помощью одного запроса:

существительное +
прилагательное в родительном падеже +

параметрическое имя в родительном падеже.

Как уже было сказано, при разработке таксономического компонента мы ориентируемся прежде всего на решение лингвистических задач: описание сочетаемости и моделей управления, исследование грамматики и семантики конструкций, деление слов на значения в словаре и др.

В то же время не лишен смысла поиск по какому-либо одному лексическому классу, причем область его применения шире, чем собственно лингвистика: ср., например, анализ упоминаний лиц женского пола в гендерных исследованиях или поиск географических названий в новостях.

3. Принципы таксономической классификации

Наша классификация базируется на принципах, заложенных в компьютерной семантической базе данных "Лексикограф: Предметные имена", созданной в ВИНИТИ РАН под руководством Е.В.Рахилиной; см. [Красильщик, Рахилина 1992]. Ключевым моментом этой классификации является то, что она производится сразу по нескольким типам параметров, т. е. является фасетной.

Известно, что грамматики давно используют многомерное разбиение на классы. Так, местоимения делятся, с одной стороны, на местоимения места, направления, времени и т. п., а с другой стороны, на указательные, вопросительные, кванторные и др. Классы каузативных и некаузативных глаголов плохо вписываются в древесную глагольную классификацию, и признак каузативности выделяется как "сквозной", пронизывающий всю глагольную систему. Критерием его выделения является то, что он предсказывает ряд особенностей грамматического поведения глаголов.

Мы пошли по тому же пути, выделяя, помимо основного семантического класса (типа "лицо" или "эмоция"), те, которые предсказывают грамматические или сочетаемостные свойства, например, класс "часть" (у имен этого класса зависимое имя в род. падеже интерпретируется как ‘целое’: *ножка стола*), класс "множество" (зависимое в род. падеже интерпретируется как ‘элемент’: *бригада рабочих*), "оценка", "семантическая одушевленность" и др.

Таким образом, каждому слову в словаре приписывается некоторый набор атрибутов разного рода:

кузов
- класс = емкость
- мереологический класс = часть
- мереологический коррелят = транспортное средство
- семантическая одушевленность = неодушевленное

интриганка
- класс = лицо
- пол = женский
- оценка = отрицательная
- семантическая одушевленность = одушевленное
- деривационный класс = nomina feminina

Заметим, что слову может быть приписано несколько атрибутов одного типа:

экспортер
- класс = лицо/организация

У многозначных слов каждому значению дается свое семантическое описание:

ножка

1
- мереологический класс = часть тела
- мереологический коррелят = лицо
- семантическая одушевленность = неодушевленное
- деривационный класс = диминутив

2
- мереологический класс = часть
- мереологический коррелят = предмет мебели
- семантическая одушевленность = неодушевленное

3
- мереологический класс = часть
- мереологический коррелят = растение
- семантическая одушевленность = неодушевленное
- ...

Как показывают приведенные примеры, параметризация значения не заменяет толкование слова, она лишь дает обобщенное описание его значения и в какой-то степени отражает семантическую структуру слова. Например, слова *ерунда* и *дребедень*, которые получили в лингвистической традиции название "оценочных

слов", охарактеризованы только по одному семантическому параметру – оценке. Напротив, у слов *интриганка* и *вертихвостка* (т. е. слов "с оценочным компонентом"), помимо оценки, заполнены и другие семантические поля.

Принцип фасетности избавляет нас от необходимости аннотировать все типы параметров. В результате этого классификация теряет единое древесное (в смысле тезауруса Роже) устройство [Кобрицов et. al. 2004]. Точнее, в классификации присутствует несколько деревьев, расположенных в разных измерениях. Например, "посуда", "предметы мебели", "инструменты" являются подклассами "артефактов"; мереологический класс "частей" имеет подкласс "частей тела", "оценка" – подкласс "положительной" и "отрицательной оценки".

Для простоты поиска и обработки информации о сочетаемости мы сознательно ограничили глубину древесной классификации. В отличие от авторов словаря [Шведова 2000], мы не стали выделять подклассы "сумчатые" или "педагоги–специалисты, воспитатели, наставники", поскольку полагаем, что они имеют мало специфических языковых черт, противопоставляющих их другим подклассам животных и людей, соответственно.

В то же время мы стремимся к тому, чтобы семантические противопоставления пронизывали разные слои лексики, например, в класс "движение" попадают не только глаголы, но и отглагольные имена типа *бег* и *шаг*; признак диминутивности приписывается именам существительным, прилагательным и наречиям; признак оценки проходит сквозь все знаменательные части речи.

На нынешнем этапе параметризованы значения порядка 40 тысяч слов знаменательных частей речи; в более полном объеме проведена разметка по деривационным признакам и аннотация имен собственных (имен, отчеств, фамилий, географических названий и т. д.).

4. Проблема "шума" при поиске и система фильтров для снятия семантической неоднозначности

На основании словарной разметки программа приписывает каждому слову в тексте кортеж параметров, имеющий структуру вида:

Номер_значения=..., ПараметрA=..., ПараметрB=... ... Параметр Z=...
Номер_значения=..., ПараметрA=..., ПараметрB=... и т.д.

Как правило, многозначные слова наиболее частотны, поэтому при поиске по заданным характеристикам исследователь нередко получает огромное количество примеров, не соответствующих первоначальному запросу (т. е. употреблений слов не в том значении, которое интересует пользователя в данный момент). Между тем, хотелось бы, чтобы пользователю были предъявлены все употребления слова, относящегося с запрашиваемому семантическому классу, и только они.

Таким образом, перед нами встает задача создать систему фильтров, которые могли бы устранить по крайней мере некоторую часть такой семантической неоднозначности. Теоретической основой алгоритмов разрешения семантической неоднозначности является контекстный анализ, широко используемый в других системах автоматической обработки текста, например, в машинном переводе, автоматическом реферировании и др.

Авторы понимают, что задача создания исчерпывающей системы снятия семантической многозначности, если она вообще выполнима, потребует колоссальных затрат времени и ресурсов, поэтому хочется отметить, что главной целью данной работы является попытка экспериментально проверить саму такую возможность. В каждом конкретном случае мы даже не будем говорить о полном снятии омонимии, но только о сокращении возможных интерпретаций (= наборов параметров), что тоже неплохо.

Созданные в ходе экспериментов фильтры используют информацию (а) о грамматических признаках контекста (падеж, одушевленность и т. д.) и (б) о сочетаемости на уровне таксономических классов. Как правило, наилучшие результаты дает применение смешанных морфо-семантических фильтров.

Правила могут быть как г л о б а л ь н ы м и – накладывающими ограничения на контекст для всех слов некоторого таксономического класса, так и л о к а л ь н ы м и – определяющими окружение данного конкретного слова (такие правила особенно характерны для тех слов, одно из значений которых употребляется в устоявшихся словосочетаниях или фразеологизмах). Во многих случаях на практике применяется комбинация таких правил.

Помимо этого, правила имеют п р и о р и т е т ы в ы п о л н е н и я . На самом деле алгоритм снятия многозначности состоит из целого ряда правил, которые применяются последовательно, сужая количество возможных семантических интерпретаций, либо непосредственно проводя выбор правильного значения. Таким образом, если какое-либо правило, на основе входных данных (исходных или полученных в результате работы предыдущих правил) выдает правильное значение, то работа алгоритма заканчивается. В результате работы алгоритма выбор правильного значения может и не состояться (в силу недостаточного описания сочетаемости или из-за принципиальной неразрешимости данной неоднозначности), в таком случае результатом работы становится сокращенное количество допустимых вариантов.

Все применяемые правила можно разделить на ограничительные и селективные. О г р а н и ч и т е л ь н ы е правила не указывают прямо на то, в каком значении употреблено рассматриваемое слово, но дают заключение о невозможности использования в данном контексте слова в одном из значений, они имеют вид:

- *слово класса X не может употребляться в контексте Y со словами класса Z.*

С е л е к т и в н ы е правила, напротив, непосредственно указывают на правильное значение и имеют вид:

- *если выполнено Условие, то слово употреблено в значении Значение.*

Отдельную проблему представляют контексты, в которых "соседи" рассматриваемого слова сами являются многозначными, и, следовательно, их нельзя напрямую использовать в качестве аргументов для разрешения многозначности. В такой ситуации следует отложить выполнение процедур снятия многозначности для данного слова и попытаться выбрать правильное значение других слов в предложении (в терминах работы [Small, Rieger 1982] выполнение текущей процедуры переходит в "спящий" режим). Впоследствии на основании новых данных можно будет совершить новую попытку выбора правильного значения существительного.

5. Пример работы морфо-семантических фильтров

Рассмотрим в качестве примера механизм работы фильтров, призванных обслуживать регулярную многозначность вида ЛИЦО – ОДЕЖДА. Слова данной группы можно разбить на подклассы в зависимости от механизма образования переносного значения. Первая группа – это слова с метонимическим сдвигом значения вида ‘одежда’ → ‘лицо’, носящий такую одежду’; сюда входят: *кокетка*, *ползунок*, *юбка*. Другая группа – это слова *амазонка, венгерка, кубанка, матроска, финка*; этот тип многозначности также можно описать как метонимию, однако направление переноса будет противоположным: ‘лицо’ → ‘одежда, ношение которой характерно для таких людей’.

Алгоритм снятия неоднозначности для слов данной группы состоит из отдельных правил, описывающих особенности употребления и свойства контекста таксономических классов *одежда* и *лицо*, составляющих эту модель.

Главным признаком рассматриваемой группы многозначных слов является разделение значений по одушевленности. Такое деление накладывает жесткое ограничение на семантические классы глаголов, при которых существительные могут выступать в роли подлежащего. Очевидно, что слова из класса *одежда*, в отличие от своих визави, не могут быть субъектом глаголов активного действия, а также вообще любых глаголов, обозначающих действия или состояния, свойственные людям и другим одушевленным объектам, ср. *Им не могут простить этого потомки тех, кого* ***амазонки били****, вселяя страх своей нечувствительностью к ранам!* (И. Ефремов).

Таким образом, на основании таксономической классификации глагольной лексики, принятой в нашем проекте, мы можем сформулировать первое ограничительное правило:

Правило 1. Слова из рассматриваемой группы в значениях, относящихся к классу *одежда*, не могут выступать в роли подлежащего при глаголах следующих классов: "эмоции", "ментальные", "физиологические", "речь", "поведение человека", "звук", "движение", "восприятие".

И наоборот, если слово является прямым дополнением к глаголу класса создания физического объекта или физического воздействия, то выбор однозначно производится в пользу значения, относящегося к классу *одежды* (Правило 2, селективное); ср. *сшить юбку/ползунки*.

Продолжая рассматривать особенности контекстов слов класса людей, мы обнаруживаем, что существует довольно обширная группа слов, обозначающих части человеческого тела и имеющих с и л ь н у ю валентность на "обладателя" этой части, которая выражается соответствующим существительным в родительном падеже. Отсюда возникает следующее селективное правило:

Правило 3. Если слово данной группы употреблено в качестве генитивного дополнения к существительному класса "часть тела", то в качестве правильного выбирается значение, относящееся к классу "лицо".

Напротив, существует группа слов *часть одежды*, с валентностью на целое. Употребление такой генитивной конструкции однозначно указывает на значения из класса *одежда* (Правило 4).

Отстранимся теперь от семантических описаний слов из нашей классификации, которые помогли сформулировать предыдущие правила выбора значения, и рассмотрим класс *людей*, так сказать, с "внеязыковой" точки зрения. Взглянем на то, какие характерные объекты действительности окружают человека в его повседневной жизни, и проверим, какие семантические описания эти объекты имеют в нашей классификации. Разумеется, невозможно описать все многообразие окружающего мира, однако, как уже говорилось во вступлении, полнота описания не является нашей задачей. Наша задача – попытаться выявить относительно простые и легко формулируемые свойства контекстного окружения слов различных классов, на основе которых можно построить алгоритмы, приводящие либо к выбору правильного значения, либо к существенному ограничению возможных интерпретаций.

Итак, человеку свойственно обладать разными вещами: это, можно сказать, неотъемлемая часть отношений человека с действительностью. Такое посессивное отношение очень часто выражается генитивной конструкцией или конструкцией с притяжательным прилагательным. Отсюда, в дополнение к Правилу 4, использующему данные об участии рассматриваемого слова в конструкции с родительным падежом, мы можем сформулировать следующее правило:

Правило 5. Если слово из рассматриваемой группы является генитивным дополнением к предметному (!) существительному, которое в свою очередь не относится к классу частей предметов, то это слово употреблено в значении соотнесенном с классом *людей*. (Такое ограничение связано с тем, что у одежды могут свои части, см. Правило 4); ср. *Ничего, кроме уздечки, не было на лошади и, кроме боевого* ***браслета амазонки****, на всаднице* (И. Ефремов).

Еще одной особенностью слов классов людей и одежды является их разная сочетаемость с прилагательными. Так, прилагательные, описывающие некоторые физические свойства объектов, не могут служить определением к словам, обозначающим *людей*. В нашей классификации к таким прилагательным относятся прилагательные физических качеств (*крепкий*), температуры (*ледяной*), цвета (*красный*) и т. д. (Правило 6), ср. *Между многими я в особенности заметил одного посетителя в* ***синей венгерке*** (В.П. Горчаков).

Наконец, заключительное селективное Правило 7 связано с морфологической одушевленностью. Здесь ситуация вполне простая – если слово рассматриваемой группы в предложении имеет форму морфологической одушевленности, то выбор правильного значения очевиден.

Литература

1. Кобрицов Б.П., Рахилина Е.В., Ляшевская О.Н. (2004). Именная классификация как лингвистическая проблема // II Международный конгресс исследователей русского языка "Русский язык: "Русский язык: исторические судьбы и современность". Москва, 18-21 марта 2004 г. Труды и материалы. М. С. 224.
2. Копотев М.В., Мустайоки А. (2003). Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети интернет // НТИ, сер.2: Информационные процессы и системы, № 6. С. 33–37.
3. Красильщик И.С., Рахилина Е.В. (1992). Предметные имена в системе "Лексикограф" // НТИ, сер.2: Информационные процессы и системы, № 9. С. 24–31.
4. Плунгян В.А. (2004). Русская морфология и русские корпуса. *См.* настоящий сборник.
5. Сичинава Д.В. (2002). К задаче создания корпусов русского языка в Интернете // НТИ, сер. 2: Информационные процессы и системы, № 8. С. 25-31.
6. Шведова Н.Ю. (2000). Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений. М.
7. Small S., Rieger C. (1982). Parsing and comprehending with word experts (a theory and its realization). In W.G. Lehnert, M.H. Ringle (eds.) Strategies for natural language processing. LEA: 89–148.