МНОГОЗНАЧНОСТЬ КАК ПРИКЛАДНАЯ ПРОБЛЕМА: ЛЕКСИКО-СЕМАНТИЧЕСКАЯ РАЗМЕТКА В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА

SEMANTIC AMBIGUITY AS AN APPLICATION-ORIENTED PROBLEM: WORD CLASS TAGGING IN THE RNC

Е.В. Рахилина *(rakhilina@gmail.com)*

Б.П. Кобрицов *(neuralman@yandex.ru)*

Г.И. Кустова

О.Н. Ляшевская *(olesar@mail.ru)*

О.Ю. Шеманаева *(shemanaeva@yandex.ru)*

ВИНИТИ РАН, Москва

Система лексико-семантической разметки корпуса рассматривается на фоне других известных семантически аннотированных корпусов, таких как корпуса, базирующиеся на семантической сети WordNet, или FrameNet. В свете практической задачи уменьшения «шума» при поиске по семантическим признакам разработчиками корпуса приняты особые соглашения, касающиеся традиционных понятий лексической семантики и лексикографии: многозначность, омонимия, порядок значений слова.

1. Введение[1]

Эта статья продолжает серию наших публикаций в «Диалоге», освещающих работу над созданием лексико-семантической разметки Национального корпуса русского языка [Кобрицов 2003, Кобрицов, Ляшевская 2004, Кобрицов и др. 2005]. На начало 2006 г. объем так называемого «Основного корпуса» (http://www.ruscorpora.ru) составил более 120 млн. словоупотреблений, из них 100 млн. – корпус современного русского языка (1950-2005 гг). Все тексты Основного корпуса содержат три вида лингвистической разметки: метатекстовую (автор, жанр текста и т.д.), грамматическую (лемма и грамматические признаки) и лексико-семантическую (разметка по лексико-семантическим группам и словообразовательным типам). В предыдущей версии корпуса семантическая разметка распространялась только на подкорпус со снятой грамматической омонимией (3,5 млн. словоупотреблений), что было явно недостаточно для проведения лексикографических, семантических и т. п. исследований. Расширение разметки на весь объем корпуса многократно усилила возможности поиска.

Теперь на первый план выходит задача повышения точности разметки и снижения уровня «шума» в результатах поиска. В нашем проекте она связана с учетом разных значений многозначных и омонимичных слов и с правильным распознаванием этих значений в тексте.

2. Корпуса с лексико-семантической разметкой

Чтобы понять масштаб задач, стоящих перед системой разрешения лексико-семантической неоднозначности, надо иметь в виду, что на сегодняшний день в мире (!) насчитывается очень небольшое число корпусов с семантической разметкой[2]. Различия в разметке и в системах автоматического разрешения неоднозначности (WSD – word-sense disambiguation) определяются, прежде всего, тем, каковы потребности пользователей конечного продукта и каким способом (и с какими затратами) разработчики собираются добиться нужного результата. От этого зависят:

1) «глубина» различения многозначности;
2) выбор словаря или лексической классификации, к которому привязана семантическая аннотация;
3) ручной *vs* автоматический способ WSD;
4) выборочная *vs* сплошная дизамбигуация.

[1] Работа подготовлена при финансовой поддержке РФФИ (грант № 05-06-80396а) и РГНФ (грант № 05-04-04130а).

[2] Здесь и далее мы будем говорить только о разметке, сопоставляющей лексеме толкование (номер значения в авторитетном толковом словаре) или указывающей место в лексической классификации (тезаурусе). Вне сферы нашего внимания останется разметка семантических ролей предикатов (PropBank и др.), анафорических связей, темпоральных отношений и т. п.

Например, если конечным результатом является правильная морфологическая разметка текста (POS-tagging[3]), которую затем можно будет использовать в системе машинного перевода, то задачи WSD ограничиваются снятием частеречной омонимии и вовсе не требуют обращения к семантической многозначности внутри одной части речи; ср. богатую традицию таких работ на материале английского языка, для которого весьма характерна конверсия из одной части речи в другую.

Пионерские работы, связанные с полноценной семантической аннотацией текстов, предполагали привязку текстовых словоупотреблений к одному из значений толкового словаря. Наиболее известен эксперимент с определением значения слова *bank* (‘берег’, ‘учреждение’ и др.) по словарю Longmans Dictionary of Contemporary English (LDOCE) [Wilks et al. 1990]. Опираясь на кластеризацию слов в LDOCE (объединение частных значений в более общие группы), группа Й.Уилкса определила значение слова в 200 предложениях. Оказалось, что точность автоматического распознавания на уровне кластеров достигала 90%, тогда как на уровне частных она составляла всего 53%. В 1994 г. Р.Брюс и Й.Вибе продемострировали проект, в котором по словарю LDOCE вручную было размечено уже 2 476 употреблений слова *interest* ‘интерес’, ‘прибыль’ и др.) [Bruce & Wiebe 1994]. Вполне естественно, что привязка семантической аннотации к индивидуальным толкованиям в словаре требовала «штучной» работы с каждым словом, а следовательно, WSD могло быть проведено только выборочно, для одного или нескольких слов (sample method).

Современные системы семантической разметки используют привязку не к словарям, а к семантическим сетям или лексическим классификациям, среди которых наиболее популярен WordNet (http://wordnet.princeton.edu/ [Fellbaum et al. 1998]), использующий разбиение на значения из словаря Oxford Avanced Learners Dictionary (OALD). Первым на его основе был размечен подкорпус Брауновского корпуса [Miller et al. 1993], содержащий 234 136 размеченных словоупотреблений, из которых 186 575 многозначны. Затем появилась система LEXAS [Ng & Lee 1996], в которой вручную были размечены 192 800 словоупотреблений, относящихся к двум сотням наиболее частотных существительных и глаголов. Корпус SemCor [Fellbaum et al. 1998], созданный в Принстонском университете, содержал 700 000 слов, 200 000 из которых (полнозначные слова) были вручную размечены по значениям WordNet 1.6, а впоследствии автоматически перекодированы в WordNet 1.7.-2.0.

Большой корпусной материал дала реализация проектов Senseval-2 и Senseval-3[4]. В первом случае было размечено в полуавтоматическом режиме (supervised method) 13 000 словоупотреблений 73 многозначных лексем, во втором – все слова подряд в корпусе из 5 тыс. слов, в кодировке WordNet 1.7.1 [Kilgarriff 2003; Mihalcea et al. 2004][5]. Как видим, среди перечисленных корпусов сплошная дизамбигуация (all-words disambiguation) была сделана только для корпуса, тестировавшегося в проекте Senseval-3, и то на небольшом объеме текстов.

Очевидно, что чем грубее семантические противопоставления, тем проще становится задача снятия семантической неоднозначности и и надежнее – ее результаты. Однако переход от толковых словарей к семантическим сетям никак не повлиял на это обстоятельство, поскольку количество синонимических групп, в которое попадает некоторое слово, напрямую соотносится с количеством значений в словаре. Лексические классификации, ведущие свое начало от онтологий, менее чувствительны к семантическим нюансам. Они различают два значения слова, только если одно из них принадлежит классу X, а другое – классу Y. Такова, в частности, классификация лексических единиц, используемая в проекте FrameNet (http://framenet.icsi.berkeley.edu/), классификация системы SenseLearner, разрабатываемой в Ланкастере [Scott Songlin Piao et al. 2005], а также таксономии, разрабатываемые для корпусов русского языка – корпуса Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ [Виноградова и др. 2001], Синтаксического корпуса [Апресян 2005] и Основного корпуса НКРЯ (настоящий проект).

Нельзя не заметить, что степень семантической неоднозначности в этих системах зависит от количества выделяемых классов. В этом отношении наиболее подробна лексическая классификация проекта FrameNet (800 классов, фрагмент таксономии показан на рис. 1). Семантическая аннотация по лексико-семантическим группам не является самоцелью проекта: на нынешнем этапе семантическая разметка применена в экспериментальном порядке к небольшому подкорпусу из 50 текстов (тексты BNC и PennTree Bank), и проводилась только вручную. Основной задачей проекта FrameNet является разметка актантной структуры глаголов и других предикатных слов, а классификации по лексико-семантическим группам отводится вспомогательная роль. Ее дробность определяется способностью слов некоторого класса становиться аргументами предиката: например, выделение класса медицинский профессий (терапевт, окулист и др.) оправдано их участием в заполнении субъектной позиции глагола cure ‘лечить’.

[3] Part-Of-Speech tagging, приписывание информации об исходной форме слова и части речи.

[4] В пилотном проекте Senseval-1 [Kilgarriff, Rosenzwejg 2000] было размечено 20 000 употреблений 35 лексем на основе лексической базы данных HECTOR [Atkins 1993], объединявшей словарь и корпус (словарные входы были созданы лексикографами «с нуля» в результате анализа 17 млн. корпуса – первой версии BNC); впоследствии была произведена перекодировка этой разметки в WordNet.

[5] В последнее время практика разметки корпусов на базе WordNet распространяется на другие западноевропейские языки, например, немецкая версия WordNet используется в одном из текущих проектов Штуттгартского университета.

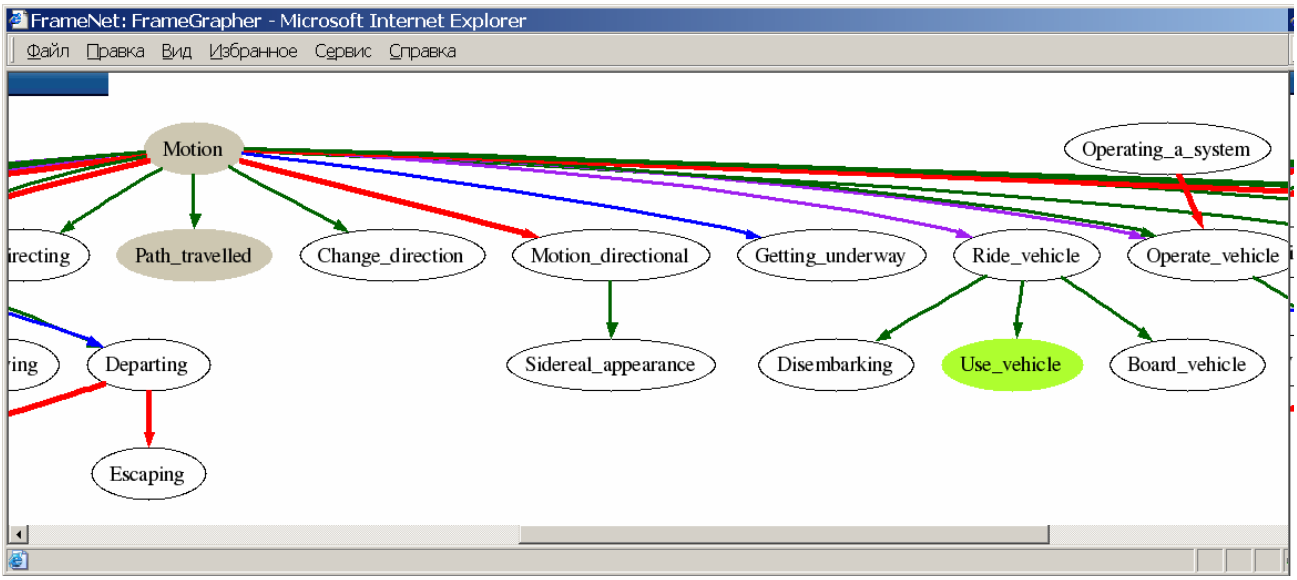

Рис. 1. Фрагмент классификации FrameNet.

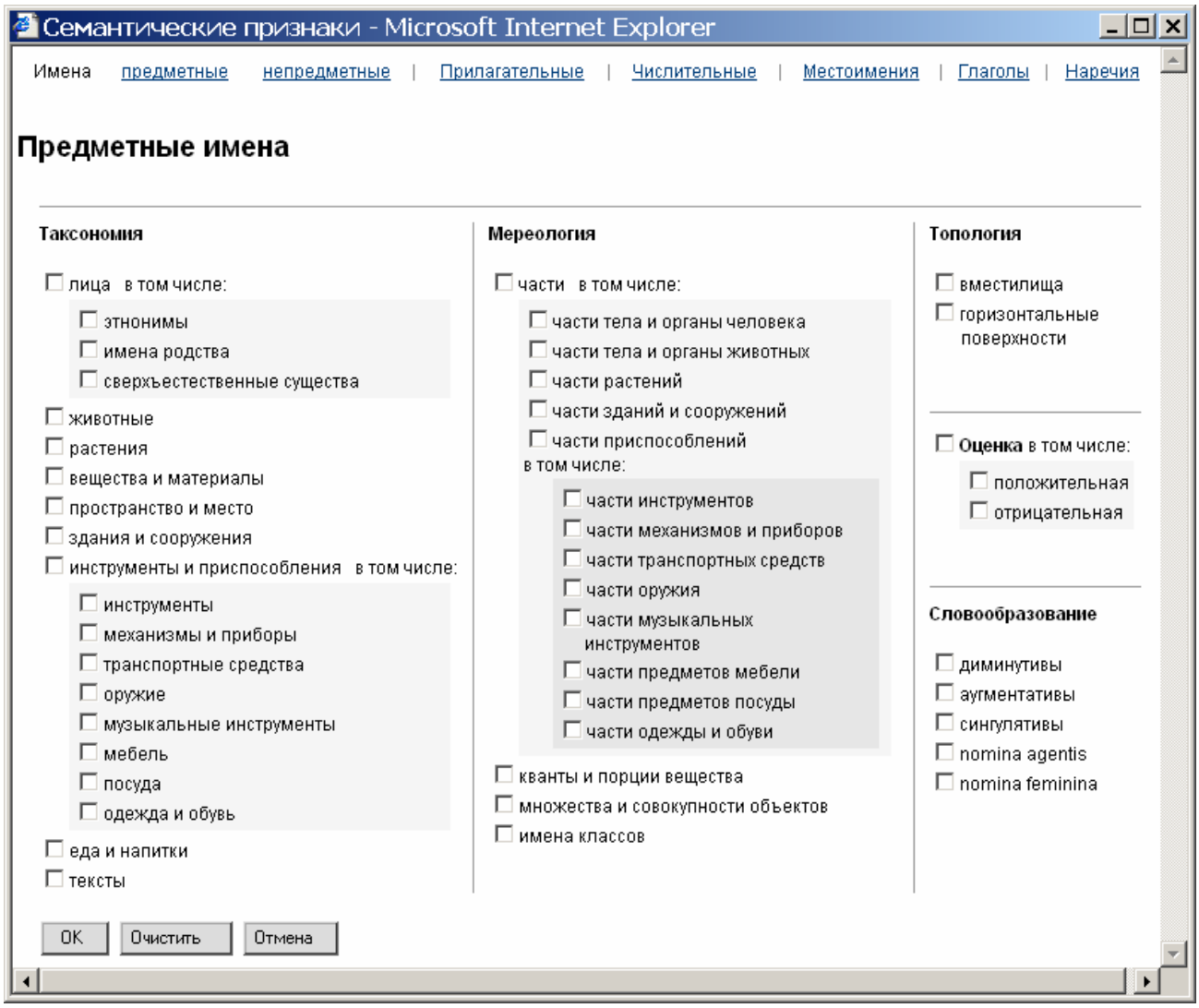

Рис. 2. Фрагмент классификации НКРЯ.

На противоложном конце шкалы – наименее детальная – классификация лаборатории UCREL (Ланкастер), которая насчитывает 232 класса (полный список классов доступен на сайте http://www.comp.lancs.ac.uk/computing/research/ucrel/usas/). Изначально исследования UCREL были связаны с

автоматическим извлечением терминологии и контент-анализом, поэтому в разных частях классификация разработана неоднородно, с большей или меньшей степенью подробности. Так, с одной стороны, в классе «средства массовой информации (media)» выделяются подклассы «книги», «газеты и др.», «телевидение, радио и кино», а с другой стороны, выделяется один общий класс «движение (moving, coming and going)».

Лексико-семантическая классификация, лежащая в основе разметки НКРЯ, по своему духу близка системе FrameNet, как по целям (обеспечение исследований лингвистов, извлечение фактов о языке), так и по происхождению (она является наследницей лексической базы данных «Лексикограф» (ВИНИТИ РАН), которая содержит форматированное толкование и информацию о модели управления глаголов в разных значениях и диатезах). Вместе с тем, наша классификация не столь детальна, как классификация FrameNet, что объясняется рядом практических соображений. Во-первых, «прямой» поиск, без построения дерева вложенных подклассов, обеспечивает быструю выдачу результатов. Во-вторых, ситуация, когда все названия семантических классов обозримы, видны в одном окне компьютера, помогает пользователю-лингвисту быстрее сориентироваться в системе классификации и, соответственно, быстро задать поисковый запрос (см. рис. 2, на котором изображена система классов предметных имен). Задача снятия семантической многозначности также оказывается проще при укрупнении лексических классов.

3. Многозначность с точки зрения лексико-семантической классификации

Итак, перед нами стоит задача разметить корпус размером в 120 млн. словоупотреблений, причем в режиме сплошной (all-words) аннотации. В идеале, неразмеченными должны остаться лишь словоупотребления, отсутствующие в словаре, а многозначные слова – получить единственно правильный разбор. Сейчас семантический словарь корпуса достиг 330 тыс. лексем (т. е. слов в одном из выделяемых значений), принадлежащих к знаменательным частям речи – именам существительным, прилагательным, наречиям, глаголам. Принципы семантической дескрипции лексем в словаре были подробно описаны в наших предыдущих публикациях [Кобрицов, Ляшевская 2004; Кустова и др. 2005], скажем только, что каждое значение слова задается набором семантических ярлыков, свидетельствующих о принадлежности лексемы к тому или иному лексическому классу, например:

парк

1) "предметное имя", "пространственный объект" (*гулять в парке*);

2) "предметное имя", "совокупность" (*парк машин*);

3) "предметное имя", "организация" (*трамвайный парк*).

валяться

1) "движение: движение субъекта", "некаузативный глагол" (*валяться в грязи*);

2) "местонахождение ", "некаузативный глагол" (*бумаги валяются на полу*).

Первичная программа семантической разметки переносит в текст наборы признаков, описывающих все значения слова; задача последующих фильтров – выбрать корректный и удалить остальные [Кобрицов 2004]. Если два словарных значения одного слова получают одинаковый набор семантических помет, например, *пломба* – ‘жестяная пластинка или сплюснутый кусочек свинца либо другого пластичного ма териала, которым опечатываются предметы, товары, помещения’ (*сорвать пломбу с опечатанной комнаты*; "предметное имя", "приспособление") и *пломба* – ‘твердеющий материал, вводимый в коронку или в полость больного зуба’ (*поставить пломбу*; "предметное имя", "приспособление"), то с точки зрения семантической разметки текста никакой неоднозначности в тексте не возникает, но – на этом уровне различения многозначности.

Соответственно, понятие многозначности формулируется иначе, чем в теоретической семантике:

Многозначность имеет место, если <u>в данной прикладной системе</u> слово описывается более чем одним набором семантических признаков

или

Многозначность имеет место, если слово входит в разные лексические классы одного типа.

(Конечно, если слово в одном из своих употреблений входит в несколько разнотипных классов, например, *молоко* – и "пища", и "жидкость", то о многозначности речи не идет).

Оказывается, что с этой точки зрения многие полисемичные слова не требуют дизамбигуации, например,

институт

1) Высшее учебное заведение;

2) Научно-исследовательское учреждение;

3) В дореволюционной России: закрытое (с пансионом) женское среднее учебное заведение для детей дворян. [Словарь Ожегова]

Все три значения описываются одинаковым образом: "предметное имя", "организация".

У имени *машина* не различаются третье и четвертое (по словарю Ожегова) значения ("предметное имя"; "транспортное средство"):

1) Механическое устройство… (*вязальная м.*)

2) Об организации… (*государственная м.*)

3) = автомобиль

4) У спортсменов: мотоцикл, велосипед.

Регулярная полисемия, с точки зрения лексической классификации, – это переход двух и более членов одного класса в другой класс[6]. Понятие регулярной полисемии важно при разработке правил снятия лексико-семантической неоднозначности. Правила, описывающие регулярные, продуктивные и частотные семантические переходы, наиболее эффективны, ср.:

(1) "размер: большой" → "степень: большая"

(2) " размер: большой" → "количество: большое"

(*большой, огромный, значительный* (1,2), *бесконечный, гигантский, безграничный, крупный, глубокий* (1), *обширный* (2)).

Считается, что решение проблемы неоднозначности в компьютерно-ориентированных системах делает также нерелевантным противопоставление омонимии и полисемии [Ravin & Leacock 2002, Kilgarriff 2003]. Представляется, что с точки зрения машины абсолютно все равно, существует ли этимологическая связь между двумя значениями имени или нет. Однако тут мы готовы поспорить.

Дело в том, что при поиске по семантическим признакам оказывается очень важным противопоставление «первое–непервое значение слова». Вероятность употребления слова в тексте в первом значении, как правило, намного выше вероятности его употребления в других значениях. Кроме того, в правилах семантической дизамбигуации лексико-семантические признаки контекста, приписанные первому значению слова, имеют гораздо больший вес. Соответственно, если слова считаются омонимами, то признак первого значения приписывается каждому из них.

4. Соглашение о первом значении, принятое в семантической разметке НКРЯ

При определении первого значения в толковых словарях лексикографы руководствуются принципом словообразовательной истории, машинный подход, напротив, руководствуется теорией вероятности: какое значение наиболее частотно и нечувствительно к контекстному окружению [Азарова и др. 2004]. Отсюда возникают конфликты между нумерацией словарей (этимологической) и реальным узусом. Ср. слово *пигалица*, имеющее следующие словарные значения:

(1) ‘птица’ – ни одного примера в НКРЯ;

(2) ‘легкомысленная девочка/девушка’ – 17 употреблений;

другие показательные примеры:

Европа (1) мифическое существо (*похищение Европы*) и (2) топоним (*посетить Европу*); *Коньяк* (1) провинция во Франции и (2) напиток; *Уран* –мифическое существо, планета, вещество.

Примеры такого рода обнаруживают определенную системность и объяснение: действительно, вероятность встретить в русском тексте упоминание мифологического персонажа или французской провинции ниже, чем вероятность обозначения бытовых или экономических реалий. В таких случаях допускается «техническая» перенумерация нумерации значения[7].

Мы также используем прием «технического» понижения статуса одного из омонимов, если он является редким словом, ср. *пара*, *сестра* и *Пара*, *Сестра* (названия рек), *сила* и *Сила* (имя), *яма* и *Яма* (бог и река); ср. также омографы *тишина* – *Ти́шина* (фамилия). Чаще всего это касается имен собственных, омонимичных нарицательным. Аналогичное решение допускается и для частичных морфологических омонимов слова, например, прилагательного *поло́вый*, обозначающего бледно-желтую масть животного и относящегося к классу прилагательных цвета. Из проанализированных нами 1000 употреблений формы генитива *полового, половой* меньше 1% имеют значение цвета (*поло́вые щенки, половые чирки*), остальные относятся к парадигмам прилагательного *половой* ‘относящийся к *полу*’ (в разных значениях) и существительного *половой* ‘слуга в трактире’. В связи с этим было принято техническое решение удалить у прилагательного *половый* в словаре признак «первое значение», но приписывать его с помощью фильтров в конструкциях *половый* + S."животное" и *половый* + S."цвет".

Итак, в корпусе НКРЯ решается задача сплошной семантической разметки очень большого объема текстов, которую можно выполнить только в автоматическом режиме. Одним из приемов в борьбе с многозначностью, которая порождает шум при поиске по семантическим признакам, становится оптимизация исходного семантического словаря, а именно, установление иерархии значений и, в случае необходимости, их перенумерация. Дополнительный критерий семантического запроса «искать только по первому значению слова» позволит обеспечить выдачу наиболее вероятного значения. Таким образом, использование порядка значений слова в разметке является простым и достаточно эффективным инструментом повышения адекватности выдачи.

[6] Вместе с тем, в корпусно-ориентированном определении регулярной многозначности снимается требование, высказанное в [Апресян 1974] – что лексемы, у которых постулируется регулярная многозначность, не должны быть синонимами.

[7] Может быть, тут было бы правильнее говорить не о первом, а об основном значении слова.

Список литературы

1. *Азарова И.В., Синопальникова А.А., Яворская М.В*. Принципы построения wordnet-тезауруса RussNet // Кобозева И.М., Нариньяни А.С., Селегей В.П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2004. М., 2004
2. *Апресян Ю.Д*. Лексическая семантика. М.. 1974.
3. *Апресян Ю.Д., Богуславский И.М., Иомдин Б.Л., Иомдин Л.Л., Санников А.В, Санников В.З., Сизов В.Г., Цинман Л.Л.* Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003-2005. М., 2005.
4. *Виноградова В.Б., Кукушкина О.В., Поликарпов А.А., Савчук С.О.* Компьютерный корпус текстов русских газет конца XX века: создание, категоризация, автоматизированный анализ языковых особенностей // "Русский язык: исторические судьбы и современность." Международный конгресс русистов-исследователей. Москва, филологический ф-т МГУ им. М.В.Ломоносова 13-16 марта 2001 г. Труды и материалы. М.: Изд-во Моск. ун-та, 2001.
5. *Кобрицов Б.П.* Морфология и синтаксис в проекте «Русский стандарт» (создание корпуса автоматически размеченных текстов) // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2003. М., 2003.
6. *Кобрицов Б.П*. Модели многозначности русской предметной лексики: глобальные и локальные правила разрешения омонимии. Автореф… канд. филол. наук. М.: РГГУ, 2004.
7. *Кобрицов Б.П., Ляшевская О.Н.* Автоматическое разрешение семантической неоднозначности в Национальном корпусе русского языка // Кобозева И.М., Нариньяни А.С., Селегей В.П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2004. М., 2004.
8. *Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю.* Поверхностные фильтры для разрешения семантической омонимии в текстовом корпусе // Кобозева И.М., Нариньяни А.С., Селегей В.П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог'2005. М., 2005.
9. *Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В.*Семантическая разметка лексики в национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. М., 2005б.
10. *Ляшевская О.Н., Плунгян В.А., Сичинава Д.В.* О морфологическом стандарте Корпуса современного русского языка // Национальный корпус русского языка: 2003-2005. М., 2005.
11. *Atkins S*. Tools for computer-aided corpus lexicography: the Hector project // Acta linguistica Hungarica 41, 1993. P. 5–72.
12. *Bruce R., Wiebe J*. Word sense disambiguation using decomposable models // Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94), LasCruces, NM, June 1994. P. 139–146.
13. *Dolan W., Vanderwende L., Richardson S*. Polysemy in a broad-coverage natural language processing system // Ravin Y., Leacock C. (eds.), Polysemy: Theoretical and Computational Approaches. N.-Y.: Oxford University Press, 2002.
14. *Fellbaum C., Grabowski J., Landes S*. Performance and confidence in a semantic annotation task // Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. Cambridge (Mass.): The MIT Press, 1998.
15. *Kilgarriff A., Rosenzweig J.* Framework and Results for English SENSEVAL. Computers and the Humanities, 34, 2000. P. 15–48. http://www.lexmasterclass.com/people/Publications/2000-KilgRosenzweig-Senseval1frame.pdf
16. *Kilgarriff A*. "I don't believe in word senses" // Nerlich B. et al. (eds.), Polysemy: Flexible Patterns of Meaning in Mind and Language. Berlin, Mouton de Gruyter, 2003.
17. *Mihalcea R., Chklovsky T., Kilgarriff A*. Framework and results for English SENSEVAL // Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, July 2004, Barcelona, Spain. 2004. P. 25–28. ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-09.pdf.
18. *Miller G., Leacock C., Randee T., Bunker R*. A semantic concordance // Proceedings of the 3rd DARPA Workshop on Human Language Technology, Plainsboro, New Jersey, 1993. P. 303–308.
19. *Ng H.T., Lee H.B*. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach // Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96), Santa Cruz, 1996.
20. *Ravin Y., Leacock K*.Polysemy: an Overview // Ravin Y., Leacock C. (eds.), Polysemy: Theoretical and Computational Approaches. N.-Y.: Oxford University Press, 2002.
21. *Scott Songlin Piao, Rayson P., Archer D., McEnery T*. Comparing and combining a semantic tagger and a statistical tool for MWE extraction // Computer Speech & Language, Vol. 19, 4 , 2005, P. 378-397.
22. *Y. Wilks, D. Fass, C. Guo, J.E. Mcdonald, T. Plate, B.M. Slator* (1990). Providing Machine Tractable Dictionary Tools, in Machine Translation, 5: 99-154.